8-3-2020

Fabrication, Testing and Theoretical Background of a Quantum Dot Floating Gate Flash Non-Volatile Memory (QDG NVM)

Nathan Butterfield
nathan.butterfield@uconn.edu

Fabrication, Testing and Theoretical Background of a Quantum Dot Floating Gate Flash Non-Volatile Memory (QDG NVM)

Nathan Robert Butterfield

BS, University of Connecticut, 2015

A Thesis

Submitted in Partial Fulfillment of the

Requirement for the Degree of

Master of Science

At the

University of Connecticut

2020

APPROVAL PAGE

Master of Science Thesis

Fabrication, Testing and Theoretical Background of a Quantum Dot Floating Gate Flash Non-Volatile Memory (QDG NVM)

Presented by

Nathan Robert Butterfield BS

Major Advisor_____

Dr. Faquir Jain

Associate Advisor_____

Dr. John Ayers

Associate Advisor_____

Dr. Helena Silva

University of Connecticut

2020

ACKNOWLEDGMENTS

I would first and foremost like to thank Dr. Faquir Jain for sharing his extensive knowledge in the field of semiconductors; through his teaching of several classes, the privilege of working in his research lab fabricating various devices and reading the publications resulting from his work. The diagrams produced in this thesis were derived from the several volumes of notes Dr. Jains produced for his students. Thank you for all the classes you have taught and guidance throughout this master's thesis research.

Thank you to Dr. Ayers and Dr. Silva for being on my advisory committee and for teaching me throughout my engineering education, in the fundamentals of electronics, semiconductors and memory technologies.

I would also like to thank Roman Mays a PhD candidate working under Dr. Jain, who tirelessly worked with me in the fabrication and testing of this QDG NVM. Thank you to all the professors and students who share a passion of electrical engineering, and through their work have contributed to the advancement of scientific knowledge and my personal growth.

Thank you to my wife Lorena who has been with me throughout my entire formal education process as well as facing life challenges together during this period. Thank you for your support, listening and love which as uplifted and motivated me toward the completion of this chapter of our life. Thank you to God, it is through the pursuits of science we may observe the elegance of nature and utilize these insights to advance humanity.

TABLE OF CONTENTS

1.0 Introduction:

Information is essential to communication, a set of directions to execute tasks, instructions to build any device or the blueprint to create biological entities. Information is ordered data, and it is conveyed through symbols, letters, and numbers. It is critical knowledge that must be preserved over time, accessed when needed, transmitted to others and processed to make useful. The space required to store information, the speed at which it is readily available, and the speed of data computation is the primary limiting factor of progress. The Metal Oxide Semiconductor Field Effect Transistor (MOSFET) is the electronic device that processes information and stores it, comprising modern cell architecture, both volatile memory and non-volatile memory (NVM). Development regarding information flow has been accelerating at an exponential rate and with the advent of algorithms and computers the pace of growth is astounding.

In 1960, Gordon Moore recognized this growth rate, predicting transistor density and processing speed would double every two years, while the cost and switching energy would reduce by half, until around the year 2020. The graphs in figure 1.1 show the reduction in power consumption throughout Moore's Law and transistor size reduction by tracking the size of a Static Random-Access Memory (SRAM) cell which consists of six transistors, Contacted Gate Pitch (CGP) which is the spacing between individual transistors and the minimum spacing between wires (M1) [1]. Although other technologies have been utilized for data storage, processing has been primarily handled by the transistor since Moore's Law was established and more recently transistors with Silicon Nitride Si_3N_4 floating gate are used for data storage as well [26][27]. The

transistor has been an ideal technology for handling data; fundamentally it is a fast switch, requiring ever less space and power to toggle between on and off.

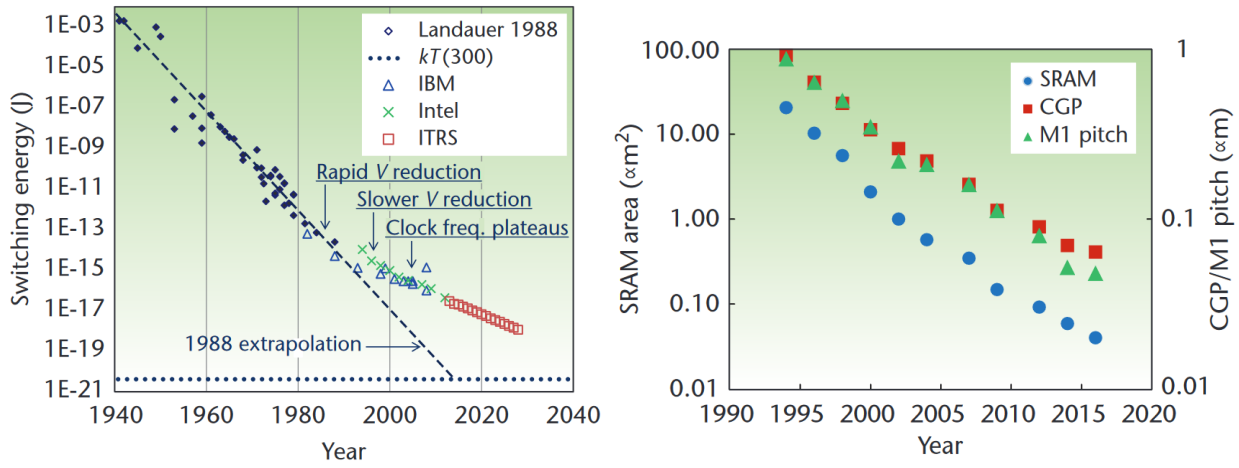


Figure 1.1: Moore's Law Switching Energy (Left) SRAM Area (Right) Over Time [1]

The essential innovation of computers is how data is stored, processed and transmitted, by breaking down any information (text, image, audio, video, etc.) into a binary equivalent using code systems such as ASCII or UCS. Binary is a two-state system of data storage, on or off, true or false, presence or absence, 1 or 0, each state is a bit and eight bits equals one byte. Early forms of binary data storage were mechanical, such as punch cards, developed in 1725 to control textile looms, which utilize a piece of stiff paper with the presence or absence of a physical hole, a 1 or 0, spaced periodically storing information. This early form of program storage was utilized in computers up until the 1980's, however the space required to store punch cards and the speed they can be processed is limited compared to magnetic tape and various transistor memory technologies. A detailed portrait of Jacquard woven into textile in 1839 required 24,000 punch cards and the fastest punch card readers had processing speeds of 30kbit/sec. Compare that to SRAM which takes 10ns to access a single cell

or speeds of 100Mbit/sec and current chip densities are achieving greater than 10 billion transistors per integrated circuit (IC).

Several data storage methods overlapped punch cards throughout the decades but the most widely used has been magnetic tape storage invented in 1928 and still used currently. Using a material that is susceptible to magnetic fields bits can be programmed into its surface. By passing a read / write head to a specific location on the tape and turning on a high intensity magnetic field the atoms of the material become polarized, namely the electrons orient themselves towards one side of each electron shell causing an overall distortion in the electron cloud. When the field is removed the material alteration is long lasting, approximately 20 years, and the polarized atoms create a magnetic field of their own, the presences or absence of it, is binary. When the read / write head is passed over the magnetic tape surface, with a lower intensity magnetic field, the programmed bits interact with the read magnetic field conveying the state of the stored data and is translated into an electrical signal for further processing. Magnetic storage such as cassette tapes and floppy disks and hard drives have been increasingly replaced by transistor-based technologies, such as Flash NVM, starting in the 1990s and currently being rendered obsolete. Flash NVM array architectures are NAND and NOR, these names are based on digital logic gates. Current magnetic tape bit densities are comparable transistors, but the technology is limited by the physics of tape rotation which requires mechanical pieces. Although transistor Solid State Drives (SSD) are on track to completely replace magnetic tape as a method of storage, magnetic fields and polarization will continue as a means of conveying binary states in future memory technologies such as MRAM [28] and Spintronics [29]. The growth of

NAND based technologies including SSD show the recent and rapid growth in millions of gigabits over time, replacing magnetic bit storage as shown in figure 1.2 [2].

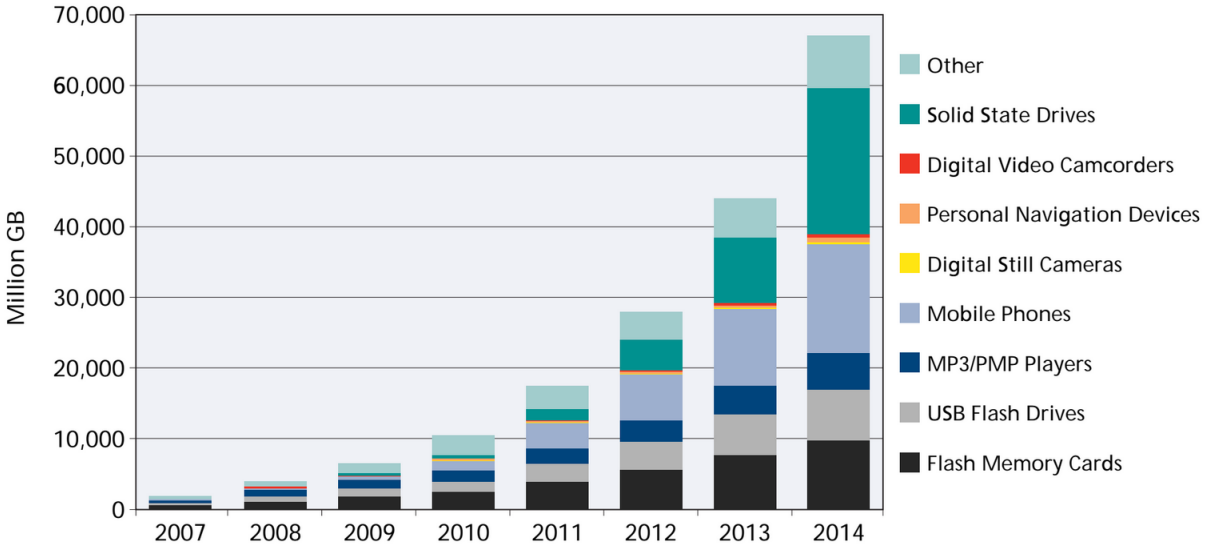


Figure 1.2: Growth of NAND Flash in Millions of GB Over Time [2]

Moore's discovery coincided with industrial production of the MOSFET however like any invention it was built of the achievements of earlier technologies. Early on the switching between binary states was handled by the vacuum tubes, invented in 1904 and still used in vintage audio equipment and high-powered transmitters. In 1925 the FET was theorized and in 1947 the Bipolar Junction Transistor (BJT) was invented leading to solid state devices replacing vacuum tube for computing purposes. It wasn't until 1959 that the first MOSFET was fabricated, although the BJT is still widely used in electronic circuits, information processing and storage is primarily controlled by the FET. As Moore's law is concluding the structure of the FET is being developed beyond the conventions of the past 50 years. All these switch variations control the flow of electrons, on or off through the device.

1.1 Semiconductor Background:

Substrates of semiconductors devices are group IV elements such as silicon (Si) and germanium (Ge) or the gallium arsenide (GaAs), which is a blend of groups III and V respectively. These substrates possess four valence electrons that form crystalline lattice structures in their intrinsic forms. This is due to the valence structure of group IV elements; electrons pair as spin up and spin down therefore a stable covalent bond between atoms is created. Each atom bonds to four other atoms forming electron pairs with no excess or shortage of electrons, forming highly ordered diamond lattices. To enhance the substrate's semiconducting properties, group III and V elements, such as Boron (B) or Phosphorous (P) correspondingly, are combined with the lattices to create an excess of charge carriers, a process known as doping. During Chemical Vapor Deposition (CVD) or Physical Vapor Deposition (PVD) a lattice undergoes high heat 1000ºC, it vibrates creating vacancies in the lattice, or intersticials, when an atomized dopant is introduced in these conditions it is incorporated into the substrate filling these voids. Group IV doped with V, contain an excess of electrons charge carriers in the lattice, are negatively charged and considered N type material, since the fifth valence electron of the dopant is unable to make a covalent bond with the otherwise stable lattice. Conversely, group IV doped with III, contains excess holes, resulting in a positively charged, P type material. The hole is an absence of an electron in the lattice, located in the valence shell of the group IV atom. These positive charge carriers conduct in E_V in the opposite direction of electron flow in E_C.

The PN junction, is either off (no bias) or operates in two mode, reverse bias (voltage applied such that current does not flow) and forward bias (voltage applied such

that current flows in one direction and is blocked in the opposite direction), refer to figure 1.1.1 [3]. The interface of a PN junction with no voltage applied experiences electron and hole migrations into the P and N type materials respectively.

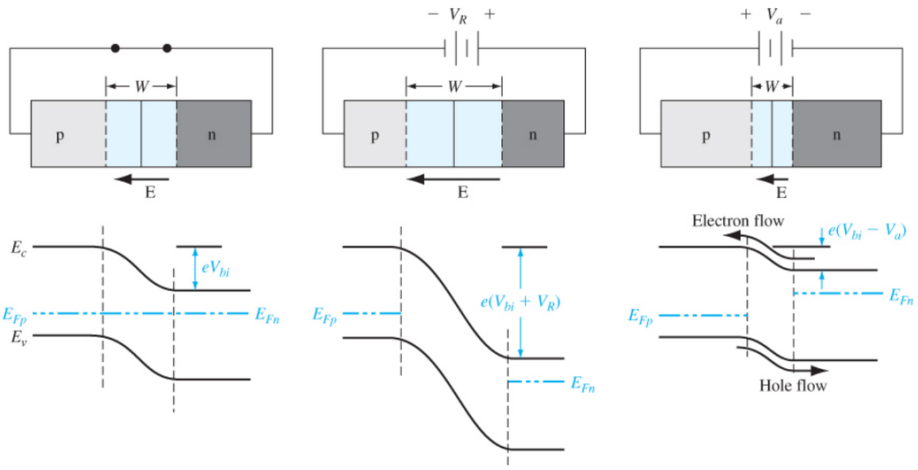


Figure 1.1.1: PN Junction Off, Reverse & Forward Bias Left to Right

This creates an uncharged barrier, void of excess charge carriers, known as the depletion region forming a stable width (W) such that charge carriers on their own do not possess the energy to transvers the barrier. P type material has a greater energy level compared to N type this results in a built-in voltage (V_{bi}) that exists across a PN junction and results in a natural bend in energy bands between P and N type materials. The Fermi level (E_F) is the probability of a charge carrier existing in its respective energy band, with the PN junction in the off state, $E_F = \frac{E_C + E_V}{2} = 50\%$ and its energy level is located at the middle of the barrier and the fermi energy for both P and N type are equal. The full probabilistic express for the fermi level is given by:

$$f(E) = \frac{1}{1 + e^{(E - E_F)/kT}}$$

Where kT is the Boltzmann Constant and temperature of the device and $f(E)$ is the probability of an electron occupying a given energy E.

The fermi level is closer to the energy band of the majority charge carrier of the material, E_V for P type and E_C for N type. Under reverse bias conditions, a voltage is applied such that the negative terminal is connected to the P type and the positive terminal is connected to the N type. The charge carrier repels away from the respective like charged terminal. This depletion of charge carriers, on both sides of the PN junction result in the increase width of the barrier and an increase in potential between P and N type energy bands, known as band bending. The flow of electrons and holes through the depletion region is blocked, as the reverse bias voltage increase so too does the depletion region width, until a point the material can no longer block current, the material fails and current rushes through the junction, this damaging state is the breakdown region.

Forward biased PN junctions allow charge carriers to conduct across the junction in their respective energy bands. With low levels of voltage applied the like charge carriers are repelled away from the terminals and toward the junction once the voltage is increase to a point to “jump” the barrier electrons and hole conducted in opposite directions. The depletion region decreases in width, the energy bands flatten, and the fermi level of the N type material is raised above the fermi level of the P type material. The voltage required to forward bias a PN junction varies depending on the built-in voltage (V_{bi}) of the semiconductor material, Si is 0.7V, Ge is 0.3V and GaAs is 1.2V, this voltage is dropped across the junction and dissipated regardless of forward voltage applied above threshold. PN junctions are present in MOSFET structures between the various regions.

2.0 MOS Capacitor and MOSFET Theory:

Metal Oxide Semiconductor (MOS) capacitor is the structure at foundation of FETs, which are used in SRAM, Dynamic Random-Access Memory (DRAM), conventional NVM, and Quantum Dot Gate (QDG)NVM. It is constructed by starting with an N or P type substrate, then growing a layer of oxide on top which is a good insulator and a thin layer of metal is deposited on top to provide an Ohmic contact. Figure 2.1 depicts the MOS structure, the substrate is grounded and the voltage (V_G) is applied to the metal.

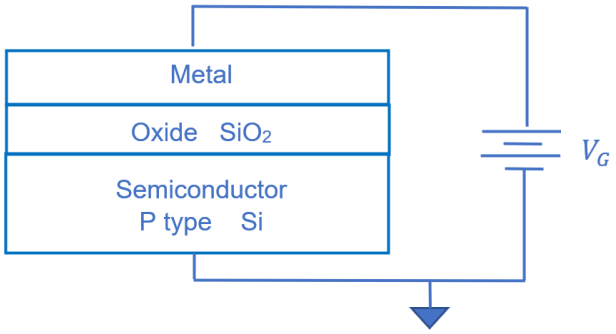


Figure 2.1: MOS Capacitor Structure

The subsequent descriptions of MOS capacitor operation pertain to P type substrates, N type substrates will operate complementary with respect to band bending directions, charge polarity and charge carriers. Figure 2.2 shows the energy band diagrams, not drawn to scale, at rest (left) and flat band (right), with the metal, oxide (SiO_2) and semiconductor (Si) layers. Energy band diagrams horizontal axis indicate spatial dimension and the vertical axis is value of energy an electron possesses with the units of electron volts (eV). At rest, no voltage applied results in the absence of charge on the metal, this layer shows a single energy level E_{FM}, indicating the close proximity of valence and conduction bands in metal. E_{FM} equals fermi level of the semiconductor,

E_F, when 0V is applied. The substrate has an even distribution of majority holes and the energy bands are bent naturally. The vacuum level (E_0) is the energy level of surrounding electrons that are not part of the material layers. The metal work function (ϕ_M) is the energy different between vacuum and E_{FM} levels, similarly the semiconductor work function (ϕ_S) is the difference between E_0 and the fermi level.

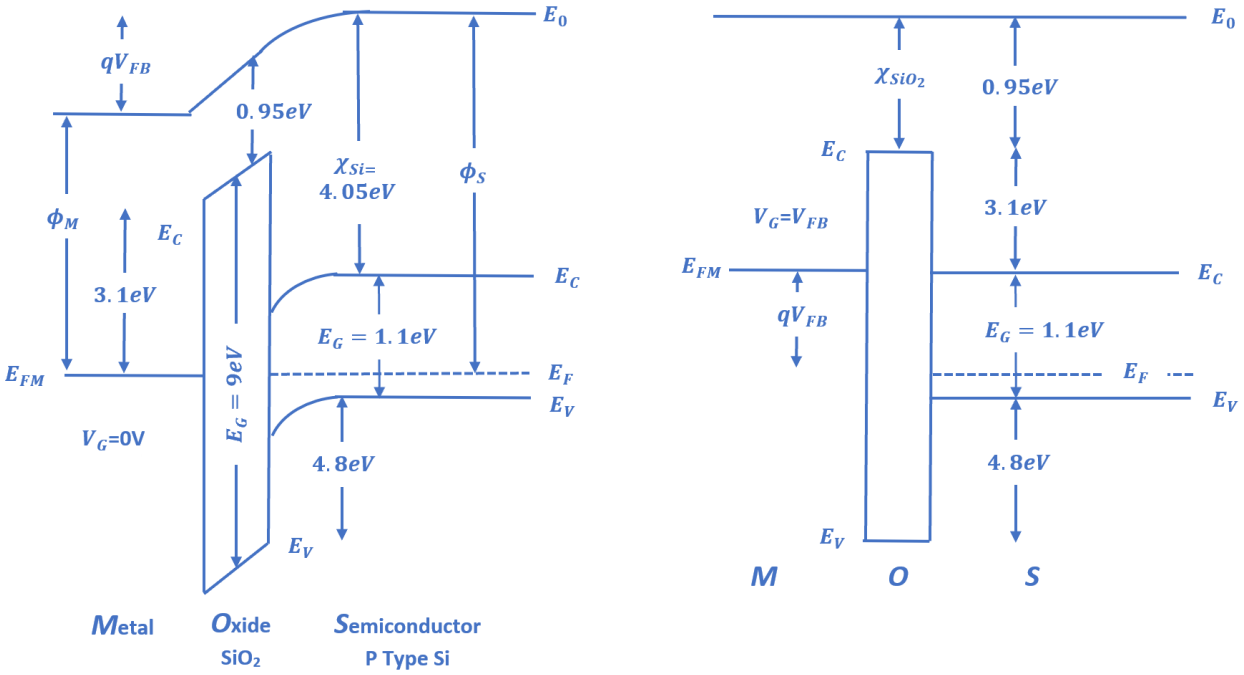


Figure 2.2: MOS Energy Bands, Rest $V_G = 0V$ (Left) and Flat Band $V_G = V_{FB}$ (Right)

Electron affinity for silicon and oxide is $\chi_{Si} = 4.05eV$ and $\chi_{SiO_2} = 0.95eV$ respectively and it is the difference between E_0 and each material's E_C. These energy levels relate to flat band voltage (V_{FB}) in the following expression, where Q_{ox} is oxide charge and C_{ox} is oxide capacitance.

$$V_{FB} = \phi_M - \phi_S - \frac{Q_{ox}}{C_{ox}}$$

When a negative voltage equal to flat band is applied to the metal, charge builds on the metal surface which raises the energy level of E_{FM} by qV_{FB} with respect to the fermi

level. This results in the bands flattening including the vacuum level $E_0 - E_{FM} = \chi_{Si}$. Flat band state is when no excess charge exists in the MOS structure. Throughout all operation regimes of a MOS capacitor, $E_G = E_C - E_V$ remains constant for each layer, namely $E_G(Si) = 1.1eV$ and $E_G(SiO_2) = 9eV$

When potential is applied to the metal contact charge builds on its surface, charge results in an electric field. The electric field passed through the insulator affecting the charge carriers in the P substrate. When a negative bias, V_G, is applied to the metal, electrons build on the surface of the metal, and E_{FM} increases in energy by qV_G with respect to the fermi level and E_{FM}>>E_C. The charge results in an electric field which repels minority charge carriers deeper into the substrate and the majority holes accumulate at the interface of the semiconductor and oxide. Figure 2.3 illustrates a MOS capacitor in accumulation, showing the physical location of charge accumulation, in addition a charge diagram is linked to the band diagram, the dashed lines link actual location of charge building in various layers, positive and negative. The bands bend lower in energy, the fermi level is lowered along with E_C and E_V below the metal level; the bending occurs in both the insulator and semiconductor as seen in figure 2.3. If voltage is decreased further the bands bend in response and the fermi level approaches E_V resulting in more holes accumulating at the interface as minority carriers are driven further into the substrate. Attempting to flow electrons through the semiconductor layer in the accumulation state would not be possible since electrons would populate holes rather than conduct. Also depicted in figure 2.3 is the voltage that builds across the oxide layer V_{OX}, in the band diagram it corresponds to the magnitude of oxide band

bending and arises due to the difference in potential between the metal and semiconductor.

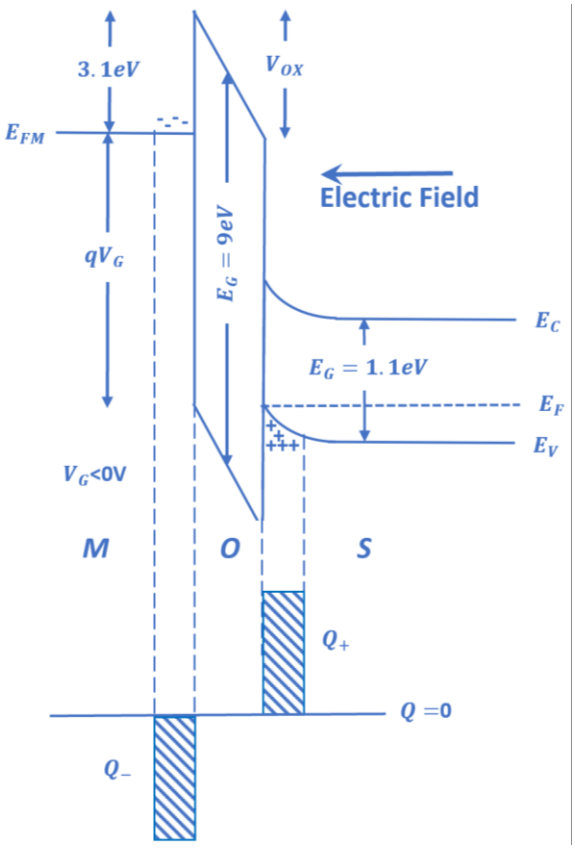


Figure 2.3: Accumulation, $V_G < 0V$ with Charge Distribution

$$V_{OX} = \mathcal{E}_{ox} * T_{ox}$$

Where $\mathcal{E}_{ox}$ is permittivity of the oxide material and T_{ox} is the thickness of the oxide layer. There exists a surface voltage (φ_s) that builds with respect to E_C not shown in diagrams.

$$V_G = V_{FB} + \varphi_s + V_{ox}$$

MOS structures have a threshold voltage (V_T), which is a potential that defines when a semiconductor is an insulator or conductor. $V_G < V_T$, including negative voltages causing accumulation, result in the semiconductor behaving as an insulator. There is a small

range of positive voltage less than V_T in which the substrate is also an insulator as it approaches conduction. This operating regime is depletion see figure 2.4 left diagram.

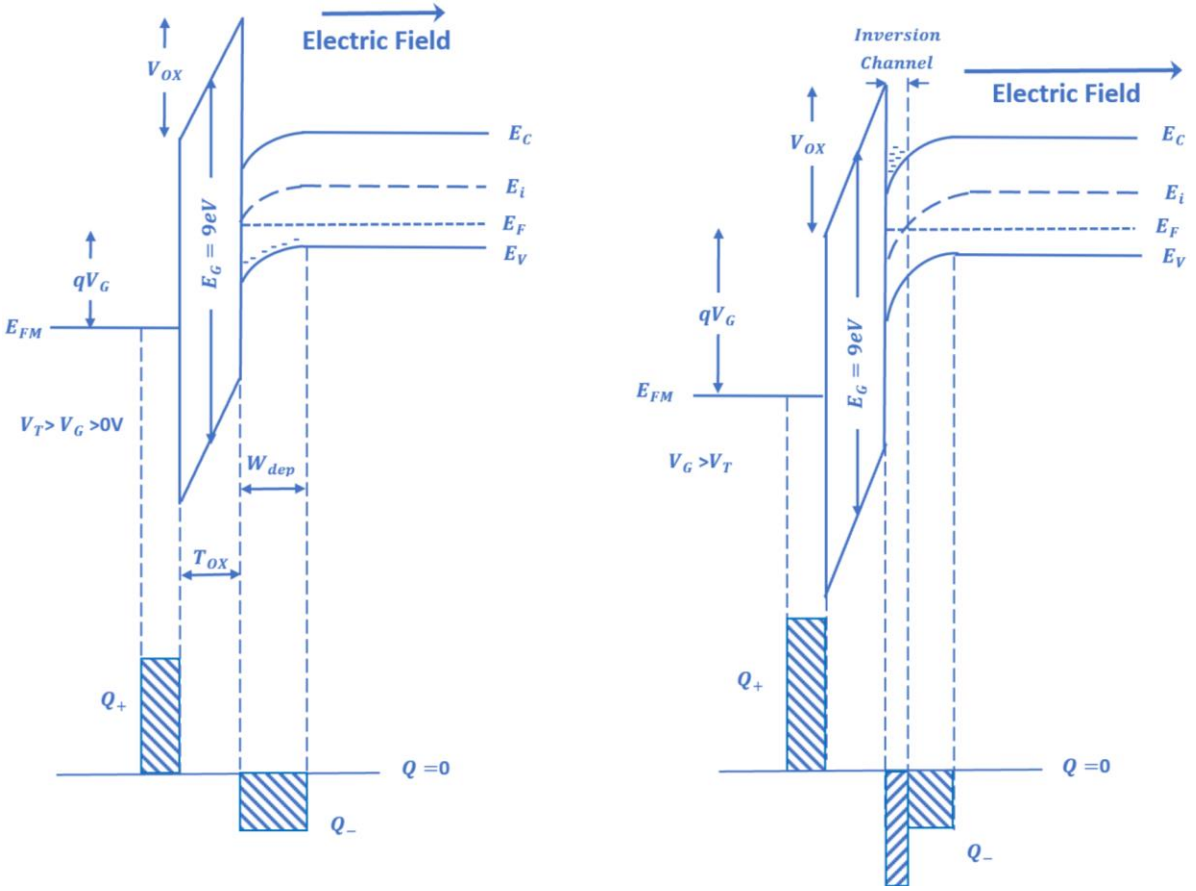


Figure 2.4: Depletion, $0V < V_G < V_T$ Left and Strong Inversion $V_G > V_T$ Right

The bands begin to bend however the fermi level remains at a lower energy compared to the bulk intrinsic fermi level E_i. As low positive voltages are applied, positive charge builds on the metal, the resulting electric field begins to attract electrons by depleting the majority charge carriers in the substrate. This is indicated by the layer of negative charge in the depletion diagram and occurs through the depletion region W_{dep}. The depletion width expression not included, is dependent on dopant concentration (N_a) and φ_s. Voltage applied to the metal can be defined with more detail utilizing the parameters of the band diagrams as follows [18]:

$$V_G = V_{FB} + \varphi_s + V_{ox} = V_{FB} + \frac{qN_a{W_{dep}}^2}{2\mathcal{E}_s} + \frac{qN_aW_{dep}}{C_{ox}}$$

where $\mathcal{E}_s$ is the permittivity of the semiconductor layer and C_{ox} is the capacitance that builds across the oxide layer. C_{ox} is defined in the following expression, where $\mathcal{E}_0$ is the permittivity of free space and A area of the gate region.

$$C_{ox} = \frac{\mathcal{E}_{OX}\mathcal{E}_0 A}{T_{ox}}$$

Once the applied voltage is greater than threshold, $V_G > V_T$, the fermi level starts to rises above the intrinsic fermi level, $E_F > E_i$, allowing additional electrons to build at the interface of the oxide and semiconductor. This is weak inversion a state in which the substrate is no longer an insulator and electrons could conduct if a bias is applied across the channel. This state is not drawn in band diagrams, however the bending and V_{ox} magnitude is between depletion and strong inversion.

As voltage increases further, charge and electric field grows larger at the metal surface, bending bands to a greater degree with electrons increasing at the interface. At a certain point saturation is reached, where charge at the interface reach a maximum, where increasing voltage ceases to alter the substrate charge carriers. This is known as strong inversion, with an external bias, electrons would also be conducting at saturation. Strong inversion is also shown in figure 2.4, right diagram. Charge carrier conduction through the MOS capacitor, under strong inversion, requires the structure to be altered by the addition of Source, Drain and a second bias applied across these contacts that bridge the inversion channel.

To create a MOSFET, the MOS capacitor is fabricated between two wells on either side of the substrate, colors in following description correspond to figure 2.5 (not

drawn to scale). These wells are doped opposite to the substrate and are known as the Source and Drain (green wells); metal contacts are deposited to facilitate biasing (purple layers). The metal contact of the MOS structure is now called the gate and the semiconductor layer (yellow) below is the channel these are separated by the gate oxide (blue layer), oxide thickness and capacitance are also defined by T_{ox} and C_{ox} respectively. The source is often grounded, and bias is applied across drain to source (V_{DS}), the resulting electric field is responsible for conducting carriers through the channel during inversion as measured by drain current (I_D). During strong inversion the layer at the interface is thin and triangularly shaped, wider toward source and narrower toward the drain, this is the inversion channel (red wedge), allowing a path for conduction of current I_D (red arrow). The electric field that arise from gate to source bias (V_{GS}), opens the inversion channel as described above. The width (W) and length (L) of the channel is indicated in figure 2.5 as well.

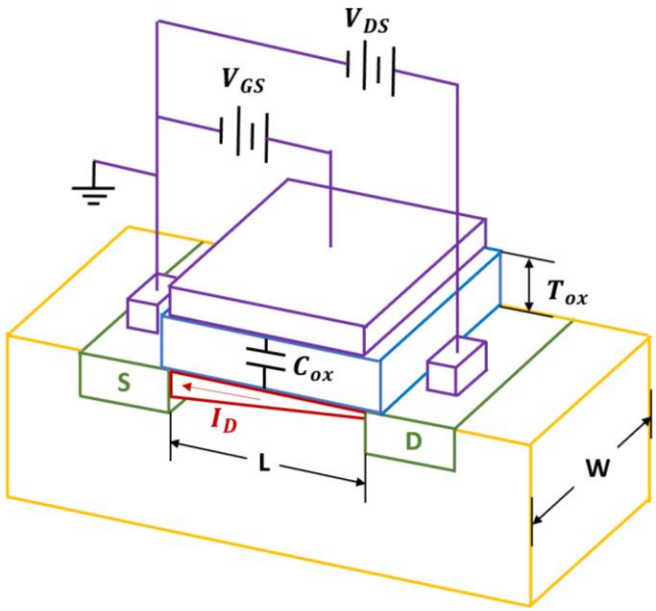


Figure 2.5: MOSFET Strong Inversion $V_{GS} > V_T$ Right

The operation states of accumulation, flat band, depletion, weak and strong inversion operate as described above, therefore the gate's electric field controls the channels ability to conduct charge carriers. Physically this is changing the shape of the inversion channel, red wedge of figure 2.5, as V_{GS} is reduced the wedge narrows, as soon as the inversion channel cease to bridge the entire channel, charge carriers stop conduction. The wedge is pinched off and the transistor is in cutoff. The inversion channel is the result of the gate's electric field driving holes deeper into the substrate and excess electrons filling the shape of the red wedge. Applying the parameters above, a set of equations define current flow through the channel, outlined in table 1.1. Table 1.1 corresponds to the $I_D - V_D$ characteristics for several V_{GS} shown in figure 2.6. Each mode of MOSFET operation requires unique expressions defining drain current.

Table 1.1: Modes of Transistor Operation, I_D Expression and Bias Conditions

Mode	Energy Band State	Equation	Condition
Cutoff	Rest, Accumulation Flat Band, Depletion	$I_D = 0$	$\boldsymbol{V_{GS} < V_T}$
Linear	Weak Inversion	$I_D = \frac{W\mu_n C_{ox}}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{{V_{DS}}^2}{2}\right]$	$\boldsymbol{V_{GS} \geq V_T}$ & $\boldsymbol{0 < V_{DS} < V_{DSat}}$
Saturation	Strong Inversion	$I_D = I_{DSat} = \frac{W\mu_n C_{ox}}{2L}[(V_{GS} - V_T)^2]$	$\boldsymbol{V_{DSat} \leq V_{DS}}$

Electron mobility (μ_n) is the ability of an electron to travel through a material for crystalline silicon μ_n=$1400\,\frac{cm^2}{V*s}$ hole mobility in the same material is slower μ_p=$450\,\frac{cm^2}{V*s}$.

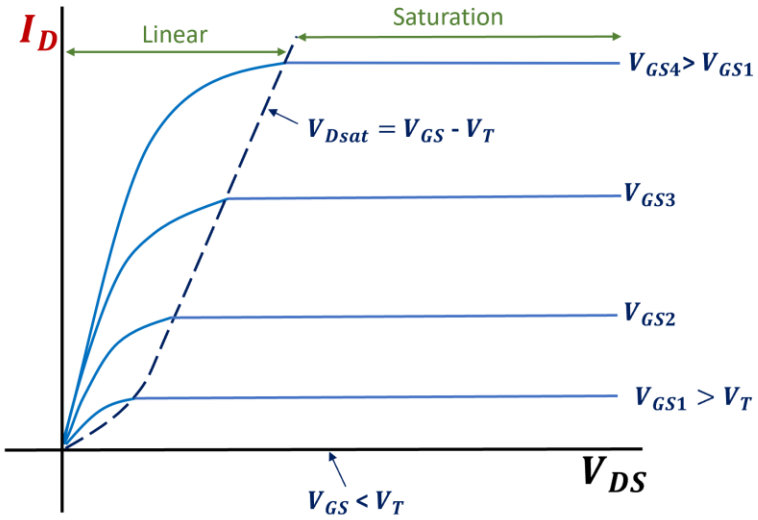


Figure 2.6: $I_D V_D$ Characteristics Increasing V_{GS} (Conduction in Inversion Channel)

There are two versions of MOSFET the Enhancement FET or NMOS which has a P type substrate, with N well source and drain. The other version is Depletion FET or PMOS which has a P type substrate a large N well surrounding the device with P wells inside of it for source and drain regions, P+. See figure 2.7 which depicts the structure of PMOS and NMOS FETs as well as their respective electronic schematic symbol. These are complementary transistors and operate in opposite manners conducting opposite charge carriers, NMOS conducts electrons with positive gate voltages and PMOS conducts holes with negative gate voltage. When combined the technology is known as CMOS (C stands for complimentary) which make the fabrication of logic gates such as the inverter possible.

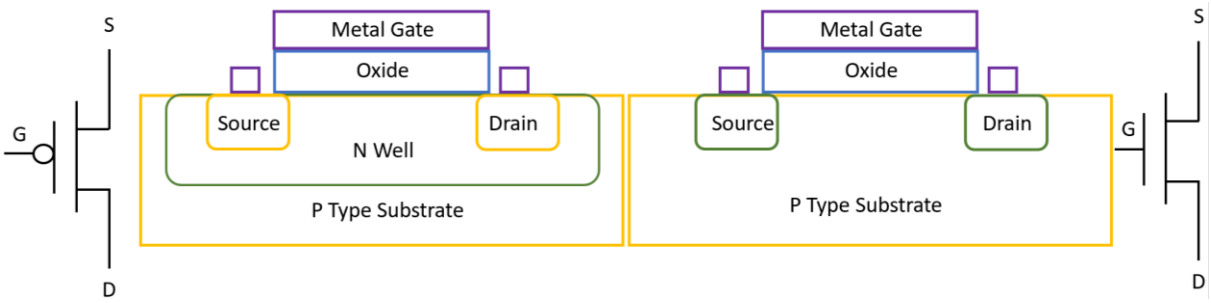


Figure 2.7: Left to Right PMOS Symbol & Structure, NMOS Structure & Symbol

The operation described above for NMOS was simulated in LTspice, see figure 2.8, to show the drain current as drain voltage is swept from no conduction of electrons to saturation. The multiple lines indicate the various gate voltages applied and the resulting impact on creation of the inversion channel, and current flow. An industry available NMOS was selected form the simulation library, V_{DS} swept from -0.1V to 1V for gate voltages V_{GS} = 0V, 0.8V, 0.95V and 1V. The complimentary operation of an industry available PMOS is simulated similarly in figure 2.9, with V_{DS} swept from -0.7V to 0.07V and complementary gate voltages compared to the NMOS simulation.

Figure 2.8: NMOS LTspice Simulation I_D V_D Characteristics

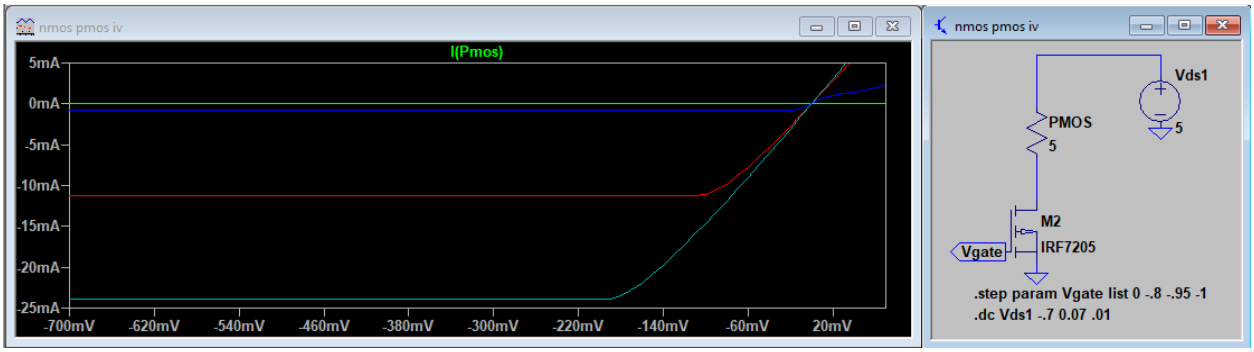


Figure 2.9: NMOS PMOS LTspice Simulation I_D V_D Characteristics

3.0 Memory Cell Theory:

MOSFETs have many uses in electronics including memory architectures for processing and storing information. Memory can be categorized as volatile or non-volatile, namely the volatility refers to memory retention once power is removed from the cell. A conventional memory cell stores a single bit, the volatile cell structures are DRAM, constructed of one MOSFET and a MOS capacitor and SRAM, comprised of 2 CMOS cross coupled inverters with 2 access FETs for a total of 6 transistors per cell. DRAMs binary information is charge stored in the MOS capacitor, while the input/output of the CMOS inverters stores the 1 or 0 for SRAM. Figure 3.1 shows how one bit is stored in a SRAM cell, its logic levels throughout the cell and which transistors are on (green) and off (red) to maintain the information as long as power (V_{DD}) is applied. Figure 3.1 also shows a DRAM cell.

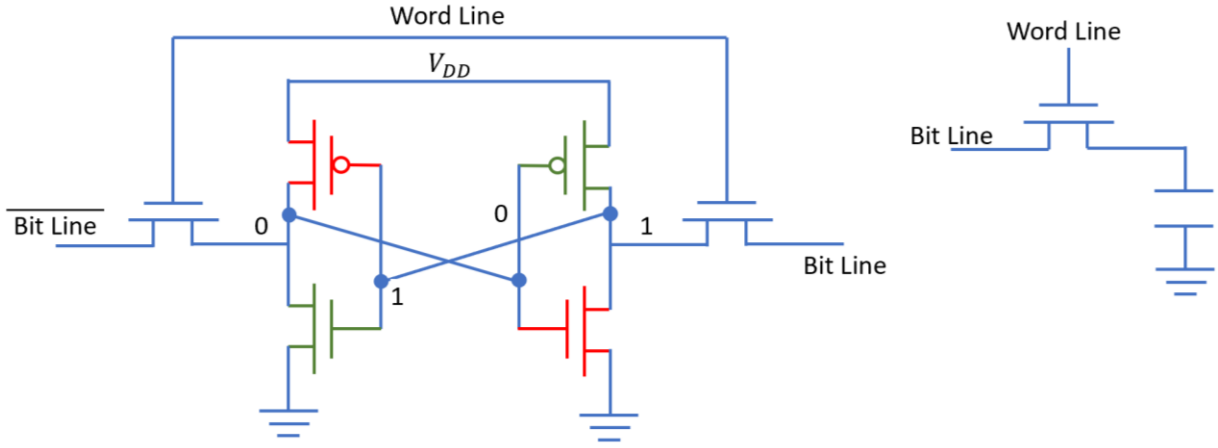


Figure 3.1: SRAM Cell (Left), DRAM Cell (Right)

When forming a memory array, with any type of volatile or NVM cell, a grid is formed of rows and columns with a single cell at each intersection. The word and bit lines allow, read write and erase operations to occur by accessing a cell or group of cells by biasing access transistors through specific row / column, known as an address.

Bit states for DRAM leak current through the MOS capacitor and requires constant checking of the state to restore charge. SRAM does not require constant refreshing of bits since the cross coupled inverters keep the information static as long as power is applied. SRAM is faster than DRAM, but DRAM has a greater bit density and is cheaper to manufacture. Both forms of volatile memory are used in modern computing however if data is not saved and power is removed that data is lost. Information therefore requires storage in non-volatile from such as Flash NVM.

NVM is required for saving information, but this memory cell is slower and requires large voltages to write and erase. If data is required out of storage it is extracted from NVM and sent to DRAM and or SRAM for processing and calculation. Flash NVM is based off the MOSFET structure with the addition of two layers, the tunnel oxide and floating gate and its operation is categorized as write, read & erase requiring various gate voltages and pulse widths. Figure 3.2 shows the structure of NVM with a conventional silicon nitride floating gate.

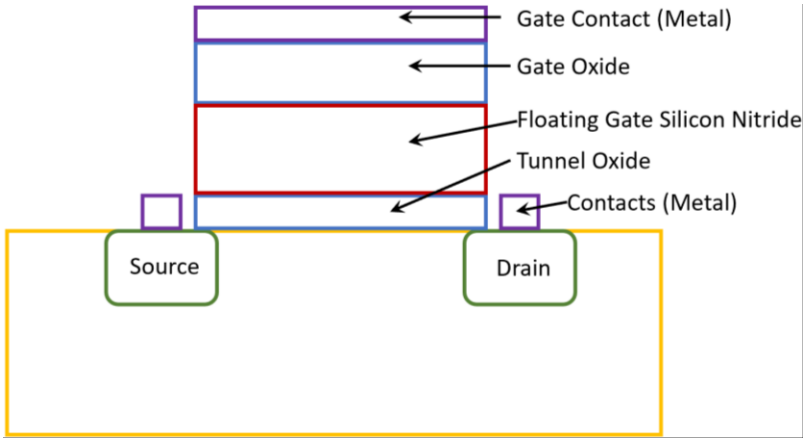


Figure 3.2: Conventional Flash NVM Silicon Nitride Floating Gate

During write operation, electrons tunnel through the tunnel oxide and into the floating gate. Conventional flash utilizes silicon nitride as a charge trapping layer, which

is an insulator that can trap electrons that receive enough energy to tunnel through the tunnel oxide. When power is removed the electrons energy level decreases and it is unable to escape the layer, trapped electrons are a stored, this negative charge is logic 0 state. Retention time for flash is guaranteed for 10 years but could last longer before charge leaks completely from these voids, through the oxide and device layers to ground.

3.1 Electron Tunneling:

Electron tunneling does not have a classical physics explanation for why these particles have the probability to cross a physical barrier such as tunnel oxide. An electron is confined by a quantum well if the well's insulation boundaries have infinite or sufficiently large potential E_C, or the physical thickness of the insulation layer T_{ox} is wide enough such that it traps the electron in the well with no probability of tunneling through the barrier. Electrons exhibit a particle / wave duality, the former is viewed as a point charge where the location is known. The wave nature of an electron is a solution to Schrödinger Wave Equation, the resulting waves are a probabilistic distribution which define the probabilities of an electron's location along that wave. Schrödinger Wave Equation is not a trivial expression, the single dimension version is included below [5]. The following figure 3.1.1 [6] shows the first solutions to the equation defining probabilities of electrons occupying space (x).

$$\frac{\hbar^2}{2m_x^*}\frac{d^2}{dx^2}\psi(x) + [E - \phi(x)]\psi(x) = 0$$

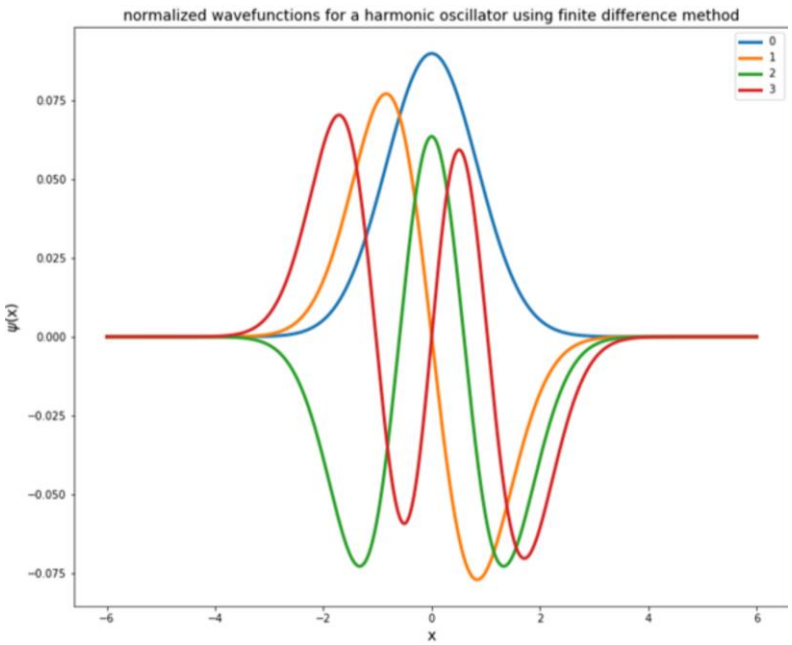


Figure 3.1.1: First Four Solutions of Schrödinger Wave Equation [6]

Where ħ is Plank's constant, E is energy, m_x^* is the electron's effective mass in the x direction, "*x is the position perpendicular to the Si–SiO₂ interface, ψ is the time-independent part of the wave function and ϕ is the potential*" ([5], page 1942). For an electron to traverse an oxide or insulation barrier it must either possess an energy level greater than the barriers E_C, or if the barrier is sufficiently thin electrons have a probability to tunnel through. Figure 3.1.2 shows the trapped energy level, in the thick-walled system (left), the electron at the trapped energy level is viewed as a particle, if viewed as a wave, the solutions to Schrödinger Wave Equation will not reach through the thick barrier. In the thin walled system (right) the electron possessing an energy less than the E_C is viewed as a wave. T_{ox} is narrow such that a portion of the wave equation solution reaches to the other side allowing the electron to tunnel. If this occurs, Schrödinger Wave Equation, is solved in three locations, the well, in the wall and the

other side of wall, the three are linearly combined producing a probabilistic solution to the wave equation that looks similar to a damped harmonic motion in figure 3.1.2, right.

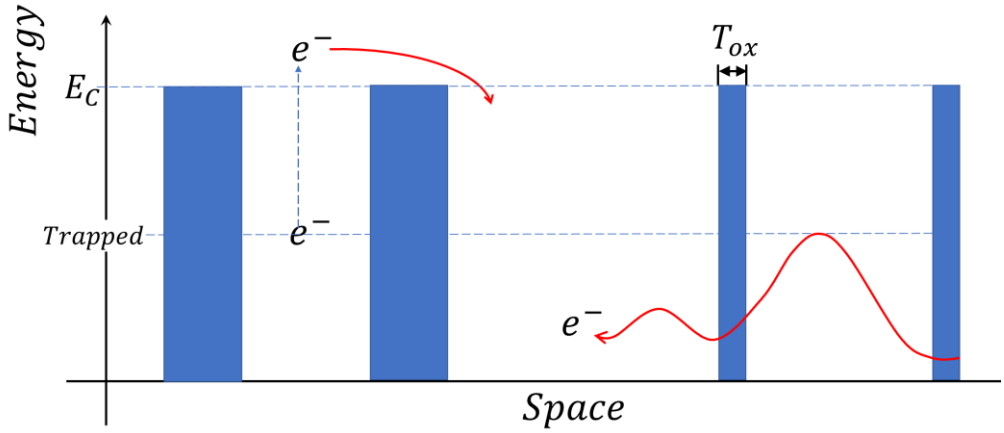


Figure 3.1.2: Electron Tunneling, High Energy (Left), Thin Wall (Right)

In conventional Flash NVM, during strong inversion electrons collect and conduct in the channel, which can be viewed as a quantum well in the sense that they are confined such that the two degrees of freedom exist. Electrons can move along L and W of the inversion channel as they conduct source to drain. Electrons are confined, above by the tunnel oxide and below by the P type substrate with excess holes absorbing electrons. However, conventional Flash NVM tunnel oxide results in the confinement of the electrons, in the channel, to be quasi bound, meaning the electrons have a probability under the right conditions to tunnel from the channel to the floating gate. The barrier near the conduction band is also narrowed by large magnitudes of band bending which permits tunneling, as well as barrier lowering which permits electrons to travel “over” the barrier, known as “lucky” tunneling electrons.

3.2 Flash NVM Operation:

The energy band diagram of silicon nitride floating gate NVM and tunneling is depicted in figure 3.2.1, left is the write operation of a logic 0, during strong inversion. Flat Band condition (figure 3.2.1 right) shows the logic 0 bit trapped and storing charge in floating gate. Once the gate voltage V_{GS} is removed, the energy levels return to the rest state, with natural band bending. The narrow barrier thickness, enough to stop electron wave tunneling and the barrier height increases stopping lucky electrons.

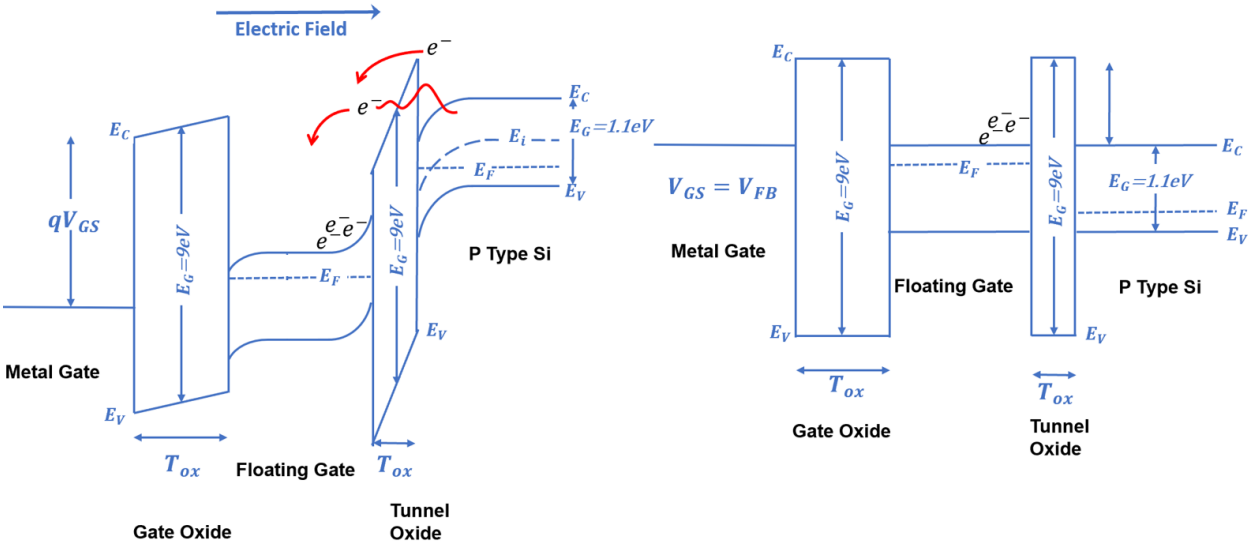


Figure 3.2.1: Electron Tunneling, High Energy (Left), Thin Wall (Right)

Two methods are used for programing states into the floating gate of conventional NVM, namely Hot Carrier Injection (HCI) and Fowler-Nordheim (FN). Charge carrier mobility for holes is slower than electrons, this makes the use of PMOS based flash NVM impractical since programming is much slower than NMOS, therefore hot carriers and charge carrier tunneling will always pertain to electrons. Figure 3.2.2 shows HCI tunneling location and NVM bias for writing a logic 0. A hot carrier is an

electron which gains kinetic energy in the inversion channel, due to less collisions, such that its energy is greater than the barriers E_C. These “lucky” electrons travel over the tunnel oxide’s E_C and are trapped in the voids of the floating gate. This is achieved by applying a bias across source and drain which causes the inversion layer to be pinched off before reaching the drain, it is at this point that electrons with kinetic energy tunnel. The pinch off is indicated by the red wedge not reaching the drain in figure 3.2.2. A relatively large, positive $V_{GS,}$ applied for a short time duration pulse train, results in an electric field that attracts electrons giving them incentive to tunnel. The E_C level for all layers are drawn alongside the NVM structure to link the physical dimensions of the device to the energy levels of each layer, the bending due to the inversion channel and electron tunneling location.

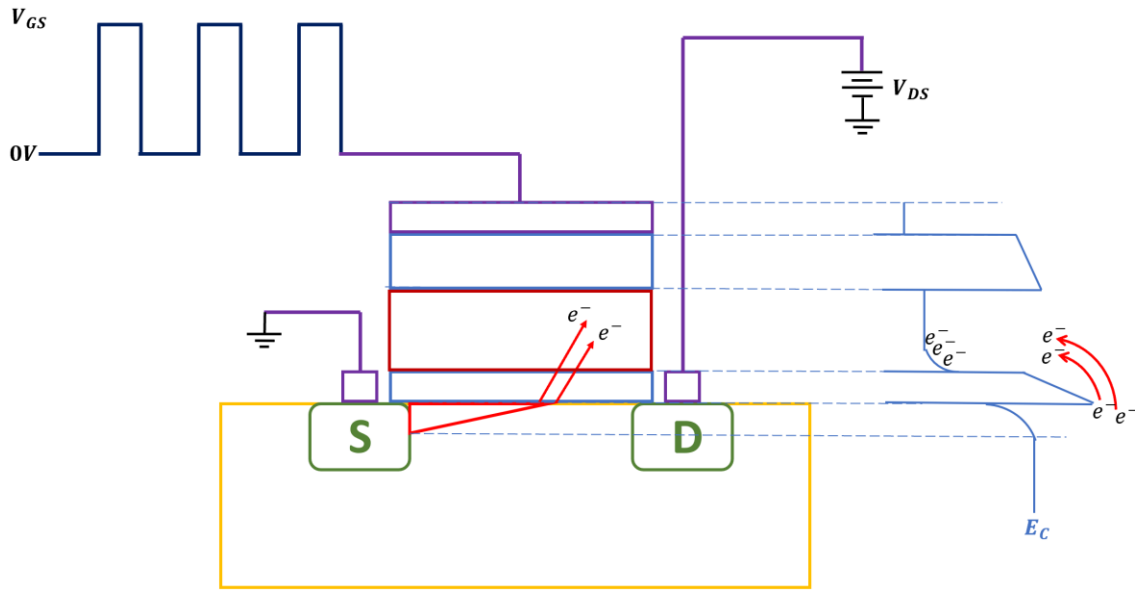


Figure 3.2.2: Writing Logic 0 to NVM, Hot Carrier Injection Method

Electrons tunneling using the Fowler-Nordheim method are helpful to view as a wave rather than a particle. The biasing of the NVM cell for FN is to ground both drain and source; therefore, electrons are not conducting in the inversion channel between

source and drain. A positive, relatively large V_{GS} voltage is applied to the gate for a single short duration pulse. This gate pulse bends the bands to a greater degree, resulting in a triangular top of the barrier's bands and a narrowing for electrons to tunnel. If a solution to Schrödinger Wave Equation, for one of these electrons in the channel, results in a portion of the damped wave extending to the other side of the oxide, at this narrow area, there exists a probability of that electron tunneling through the tunnel oxide. Figure 3.2.3 shows tunneling under FN method, band diagram, device structure and biasing for writing a logic 0 to the silicon nitride floating gate

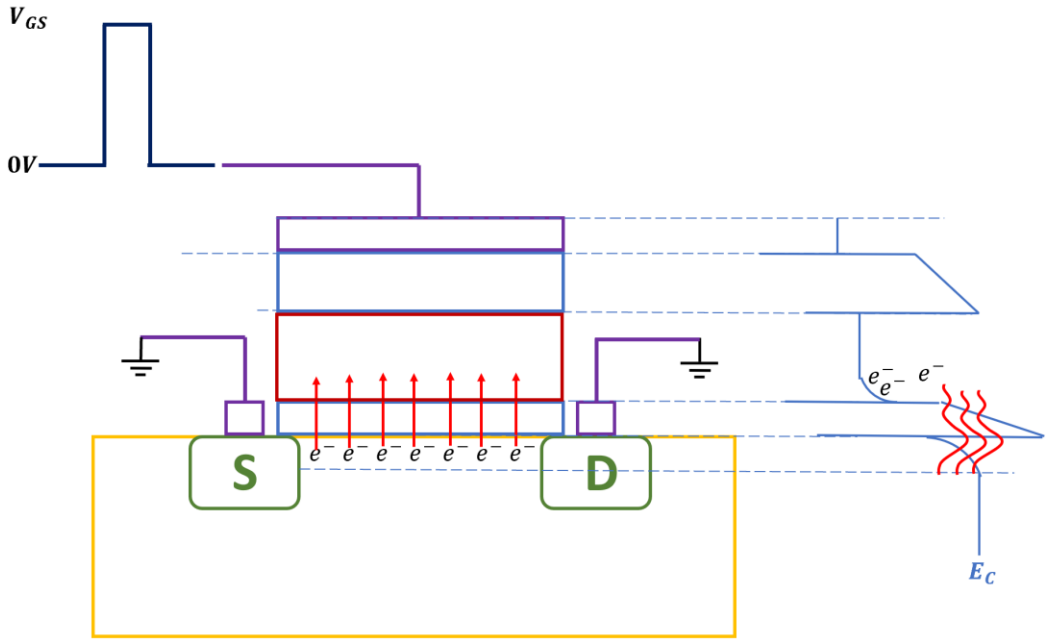


Figure 3.2.3: Writing Logic 0 to NVM, Fowler-Nordheim Method

Tunneling under FN is uniform along the channel length compared to HCI which is tunneling at the pinch off point, this means the current density is grater for Hot Carrier Injection since the area of tunneling is focused. HCI is faster to write, consumes more power and reduces the life of an NVM cell due to increased damage from higher current density. FN in contrast has slower write speeds, less power consumption and increased

endurance due to the less damage to the tunnel oxide, these properties are due to the uniform tunnel distribution and reduced current density.

The two architectures for Flash NVM are NOR and NAND, see figure 3.3.4 [2], and writing to these cells utilize the HCI and FN tunneling methods respectively. NOR has two advantages over NAND, it is faster due to HCI tunnel write operation and NOR allows the erase operation to be performed for an individual cell. NAND cells must be erased in blocks, and have slower write times due to FN, however they have one advantage over NOR, which is the bit storage density is increased which is why NAND Flash NVM architecture dominates the industry. Both NAND and NOR cells can be written individually via their respective tunneling methods, HCI and FN, but both architectures are erased using Fowler-Nordheim method of tunneling.

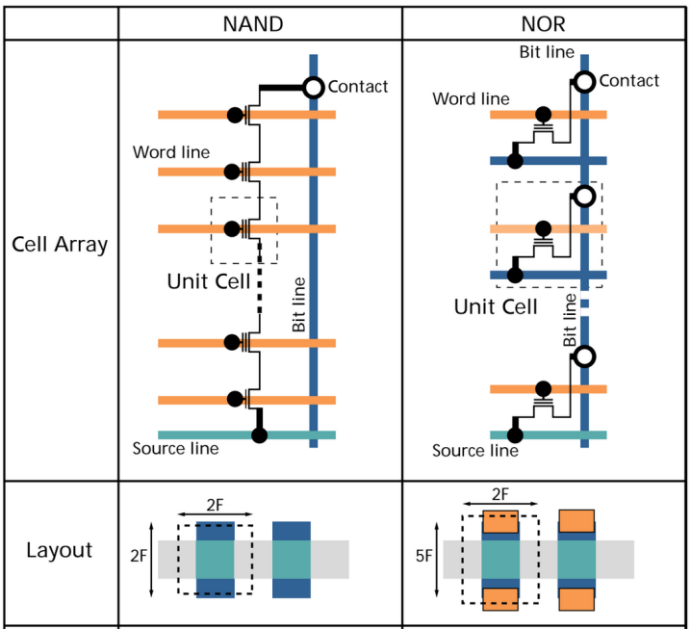


Figure 3.2.4: Flash NVM, NAND and NOR Cell Array and IC Footprint [2]

Erasing the charge from the floating by tunneling electrons back through the tunnel oxide and into the channel using FN is inherently slower than the write operation.

This is due to the lower hole mobility μ_p compared to electron mobility μ_n as outlined previously. As electrons vacate the silicon nitride, they leave holes behind, resulting in a positively charged floating gate or logic 1. The source and drain are grounded during erase operation like the FN writes operation. To create an electric field that will repel electrons out of the floating gate, a large negative V_{GS} pulse is required. The pulse width for write in FN is between 10ns to 10µs, however since erasing takes long the pulse width is increased to 10ms. Erasing a written state to logic 1 for both NAND and NOR is shown in figure 3.2.5 as electrons tunnel out of the floating gate, through the tunnel oxide, to the inversion channel and ground.

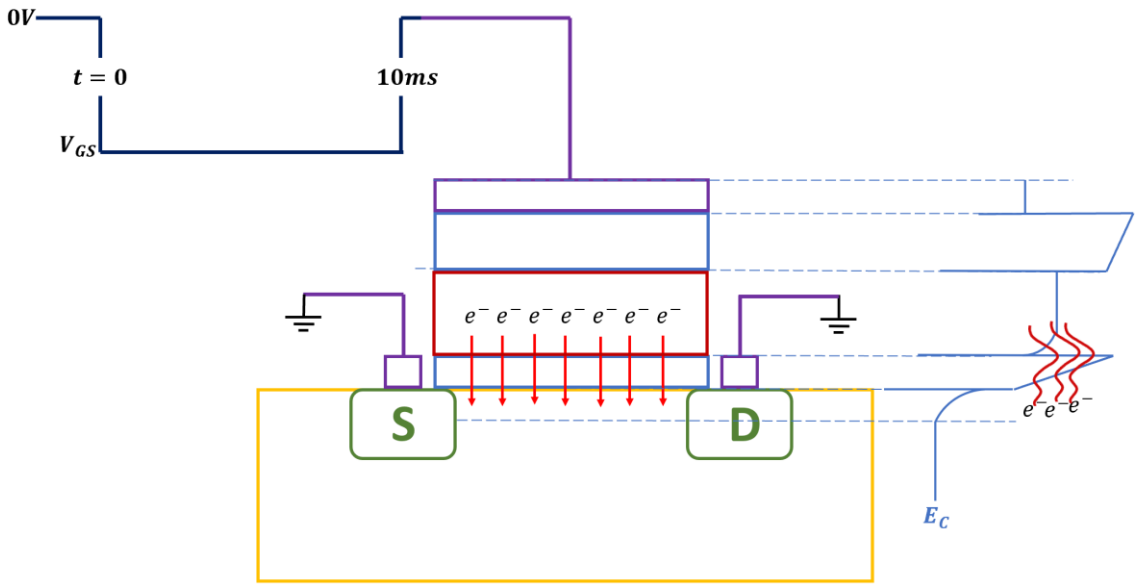


Figure 3.2.5: Erasing NAND and NOR, Fowler-Nordheim Method, Logic 1

The final operation of Flash NVM is read, to determine the informational state stored in the memory cell. If electrons are stored in the floating gate, indicating a logic 0, it results in electric field of its own. If voltage is applied to the control gate the positive charge building on the gate creates an electric field and the floating gate electric field opposes this direction, resulting in a reduced impact on the inversion channel. The

stored charge in the floating gate Q_{FG} results in a floating gate voltage with respect to ground (V_{FG}). The floating gate voltage relationship to V_{GS} and the capacitive model is given by the following expression [7].

$$V_{FG} = V_{GS}\frac{C_{FG}}{C_{FG} + C_{FB}}$$

Where C_{FG} is the capacitance that builds between the control and floating gate and C_{FB} is capacitance between the floating gate and body of the substrate, along the inversion channel. Since the floating gate overlaps source and drain, C_{ox} can be broken down into several parasitic capacitances in parallel, the primary three, are depicted in figure 3.2.6.

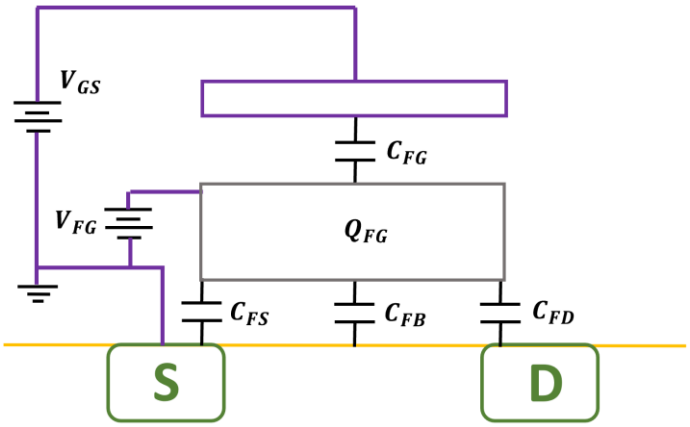


Figure 3.2.6: Capacitance Model of Flash NVM

The reduce electric field, due to the stored negative charge of the electron $-Q_{FG}$, effects the ability to open the inversion channel and thus altering V_T. This change in threshold voltage (ΔV_T) is also defined by

$$\Delta V_T = -\frac{Q_{FG}}{C_{FG}}$$

Therefore, if a logic 0 is stored, V_T is shifted higher due to $-Q_{FG}$ compared to logic 1 allowing for a read voltage to be applied to the gate to determine the bit state without impacting the information.

The voltage pulse applied to the gate during read operation (V_{Read}) is much lower than write voltages, therefore the band bending is less extreme, not allowing electrons to tunnel, preserving the information. With a read voltage pulse, $V_{T1} < V_{Read} < V_{T0}$, external sense circuitry is utilized to detect drain current, I_{DS} corresponds to the present or absent of charge in the floating gate. The NVM will turn on and conduct if no $-Q_{FG}$ is present, the floating gate void of electrons is logic 1, or the NVM will remain off if $-Q_{FG}$ is trapped in the floating gate altering the electric field preventing the inversion channel from forming, this is a logic 0. Figure 3.2.7 [8] shows the Flash NVM cell in read operation and the external current sense circuit along with the $I_{DS}V_{GS}$ characteristics, the ΔV_T shift and proper V_{Read} required to read the cell.

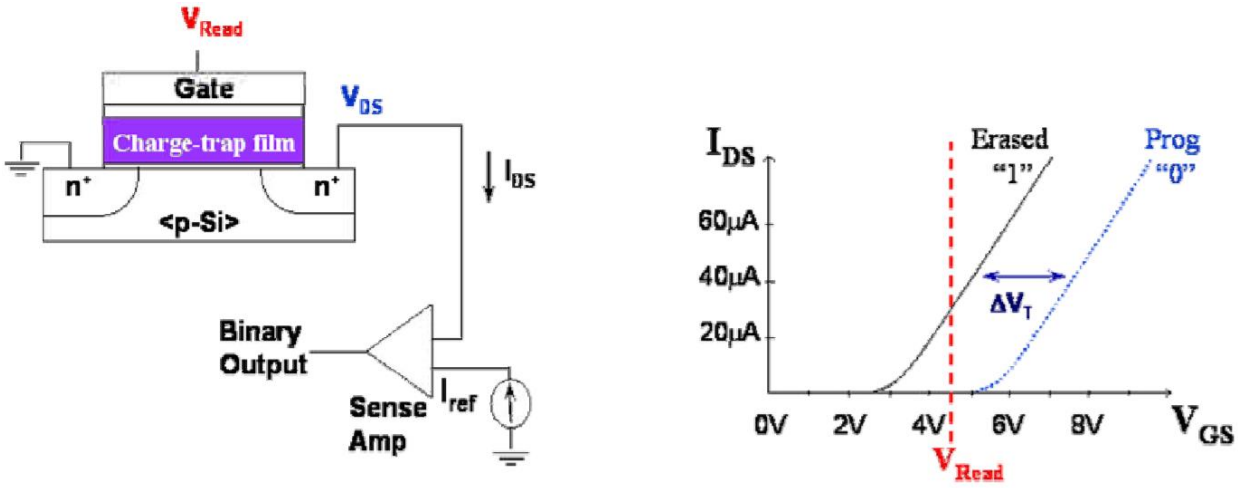


Figure 3.2.7: Flash NVM Read Operation Charge trap film is silicon nitride in SONOS (Si-Oxide-nitride-oxide-polySi) [8]

4.0 QDG Flash NVM Theory

The miniaturization and power reduction of FETs is approaching the limits of physics. Specifically, as the gate length and threshold voltages decrease controlling drain to source leakage, in the form of undesired electron tunneling through the channel, is becoming more difficult to achieve. The channel length limit has been met at 7nm. In research settings, 1nm carbon nano tube gate transistors [9] and atomic base transistors are being claimed although it is unclear if these will be practical for industry fabrication. What is clear is that Moore's Law has concluded for conventional FET and is being extended over the next 5 or 10 years via structural alterations. The FET as a device will need to be replaced with a practical technology which allows for progress to continue in storage density, processing and transmission speeds. The memory market is about $100 billion per year and growing alongside the demand for computing power, figure 4.1 [10] shows the market share of memory technologies, with DRAM dominating the volatile computational needs (63% of all memory) and NAND Flash dominating the non-volatile storage (36% of all memory). For progress to continue, the demand of the memory market must be met with novel solutions to address the end of Moore's Law.

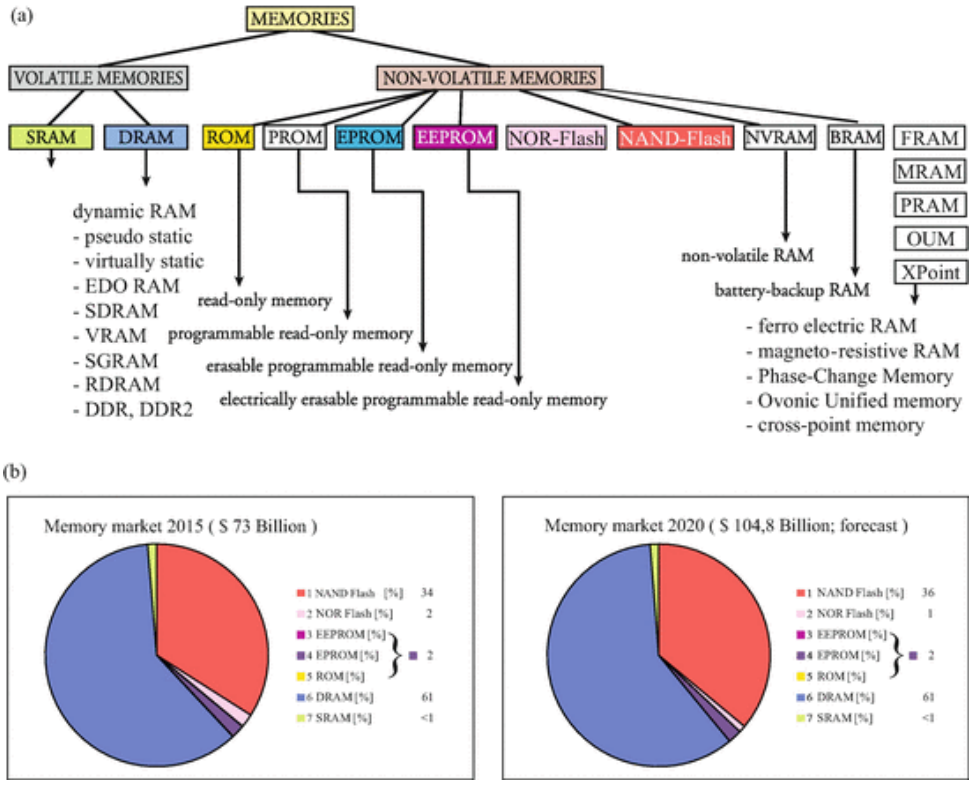


Figure 4.1: Current Memory Market [10]

The current state of FET advancement is branching in multiple directions as a technological bridge between the past and future is constructed. Some paths seek to extend the current binary FET gate miniaturization by overcoming leakage issues, via incorporation of new materials, altering conventional geometries or a combination of the two. Several paths seek to move beyond semiconductor materials and redesign the FET and transistor memory cells, such as atomic / molecular FETs, Spintronics, Nitrogen Vacancy memory cells, plasmonics, MRAM and quantum computing to name a few. Another approach is to fabricate in the vertical direction creating 3D structures allowing for increased bit density such as vertical NAND and FinFETs. An alternative to increasing bit density is to compute and store bits beyond 1 and 0, with a single device able to differentiate between multiple states.

One method to achieve multistate devices is to incorporate Quantum Dots (QD) in existing memory architectures such as Flash, SRAM, DRAM. These additional states are intermediate voltages between logic 0 (0V) and logic 1 or supply voltage of the chip (V_{DD}), often 5V but it depends on the manufacturer. The fast switching nature of the FET results in a short transition time between states, therefore intermediate voltages can only exist briefly during voltage ramp. Therefore, controlling charge storage in floating gate is binary, replacing the conventional floating gate with multilayer QD will allow the magnitude of charge storage to be more finely regulated to produce discrete intermediate states.

Quantum dots are widely used in semiconductor displays, biomedical applications and research regarding their incorporation into semiconductor devices to enhance operation. The QD structure will confine a particle to a point, restricting motion in all three dimensions, if its size approaches the de Broglie wavelength of that particle. The electron is the particle of interest to confine and it's de Broglie wavelength is 1.23nm. A physical QD structure is spherical, comprised of an intrinsic semiconductor core with a cladding layer of oxide, Si core (4nm diameter) SiO_2 cladding (1nm thick shell) and Ge core (3nm diameter) with GeO_2 cladding (0.5nm thick shell). Electrons tunnel through the cladding layer becoming trapped in the core. Germanium (Ge) has higher carrier mobilities compared to silicon with $\mu_n = 3900 \frac{cm^2}{V*s}$ and $\mu_p = 1900 \frac{cm^2}{V*s}$ for electrons and holes respectively. This is due to reduced band gap energies compared to silicon, for Ge ($E_G = 0.64eV$) and Germanium oxide GeO_2 ($E_G = 6eV$) [19]. Figure 4.2 compares the QD structure and band gaps for Si and Ge varieties.

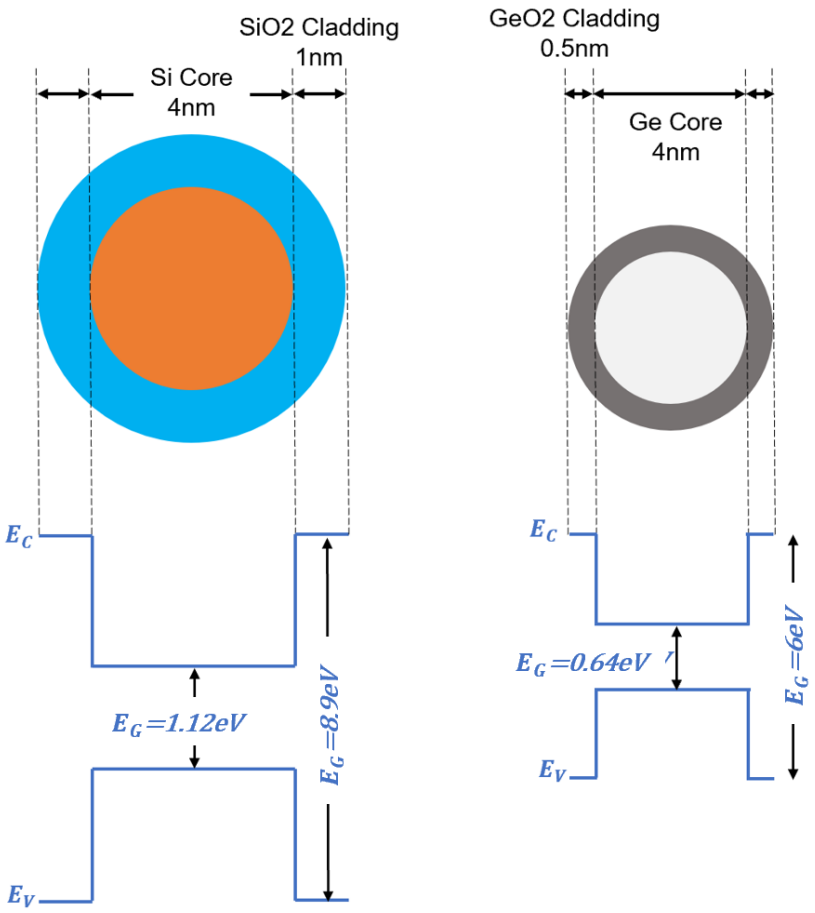


Figure 4.2: Silicon vs Germanium Quantum Dots

When QDs are assembled such that many adjacent dots form a plane, with their cladding touching one another, a QD Super Lattice QD(SL) is formed with unique properties. Confined electrons are best analyzed as waves, and a QDSL populated with electrons, experience an interaction of waves. Solutions to the Schrödinger Wave Equation for adjacent electrons in QDs have overlapping waves, this allows the electrons to share information and give rise to additional band gaps inside the well of the QD known as minibands. Minibands extend throughout the QDSL and provide discrete energy states and a free path of conduction for electrons traveling within a minibands between QD. The minibands for a Ge QD super lattice is depicted below in figure 4.3 indicated by the red horizonal lines in the QDSL energy band diagram.

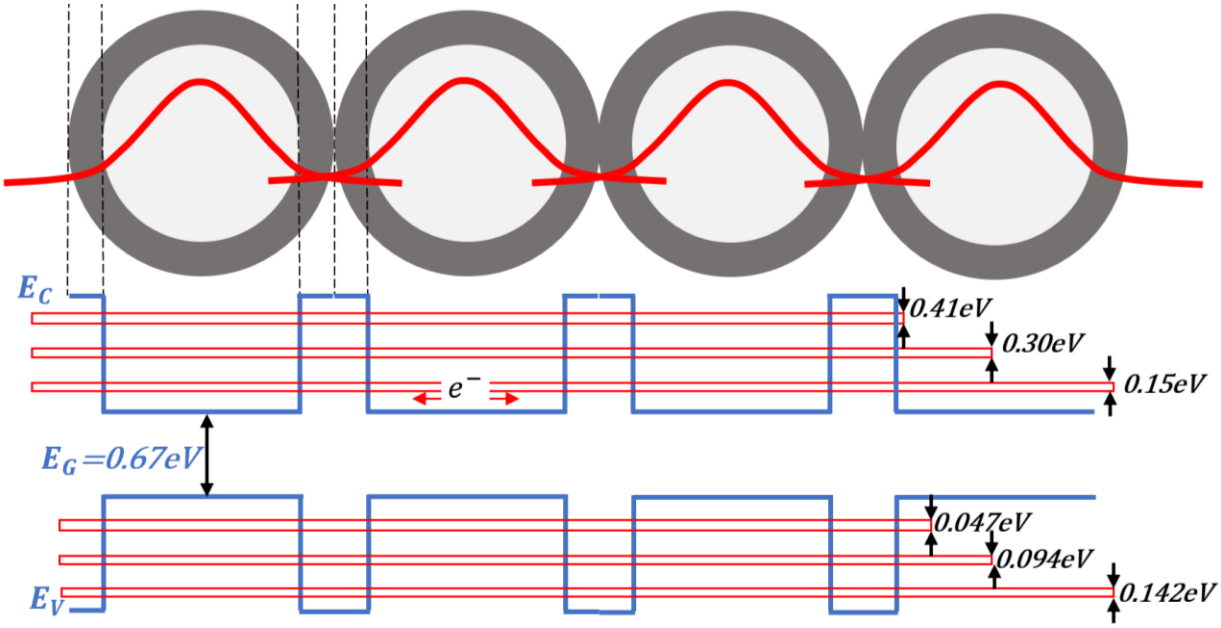


Figure 4.3: Germanium QDSL, Miniband Energy Level [19]

The first solution to Schrödinger Wave Equation populates and travels in miniband 1, which has its own $E_G = 0.15eV$. The electrons are shown in the QD core of figure 4.3 as waves and since the overlapping of adjacent electrons occurs the wave does not decay as compared to the tunneling diagram of figure 3.1.2. The second and third solutions to Schrödinger Wave Equation fill and conduct in minibands $E_G = 0.30eV$ and $E_G = 0.41eV$ respectively.

Replacing silicon nitride floating gate with a QD floating gate, comprised of multiple layers of dots, or mixed dots, will increase the number states due to the discrete minibands. The difference in programming the QDGNVM is the magnitude of write pulse voltage varies, the minimum voltage pulse required to store charge will populate the first miniband in QD layer 1. Increasing the voltage for write pulse will populate additional minibands at higher energy level, thereby allowing for the selection of QD layer and the specific miniband to store charge. When the QDGNVM cell receives a read pulse, the multistates manifest as a range of stored charge possibilities. This

corresponds to multiple ΔV_T shifts and as such discrete currents can be detected with sense circuitry and observed in I_D V_{GS} and I_D V_{DS}characteristics.

In the device fabricated for this research it was desired to have a tunnel oxide that exhibited direct tunneling. Direct tunnel occurs in thin barriers 60Å > T_{ox} >20Å. The probability of electron tunneling under direct tunneling can be defined by the following approximation [5], where x_1 and x_2 are the classical turning points, $k(x)$ is the wave number in the x direction.

$$P \approx exp\left[\int_{x_1}^{x_2} k(x)dx\right]$$

Figure 4.4 shows the $\phi(x)$ potential, first solution to the Schrödinger Wave Equation and the energy diagram explanation for direct tunneling through a barrier T_{ox} thick, with an energy E_C high, redrawn from [5].

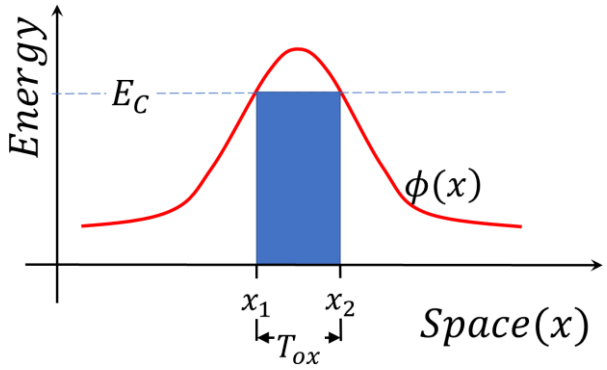


Figure 4.4: Direct Tunneling

The energy band diagram for the device fabricated during this research is shown in figure 4.5, the structure is as follows: P-Type silicon substrate, n+ source / drain, T_{ox}= 40Å for the SiO_2 tunnel oxide, 2 layers of Ge QD for the floating gate and T_{ox}= 75Å for the Silicon Nitride (Si_3N_4) control gate oxide and 1500Å of Aluminum for gate, source and drain contacts. The resulting minibands introduce intermediate states beyond the

conventional Flash NVM labeled as states 01, 10 and 11, supporting three solutions to the Wave Equation for electrons to tunnel and program states.

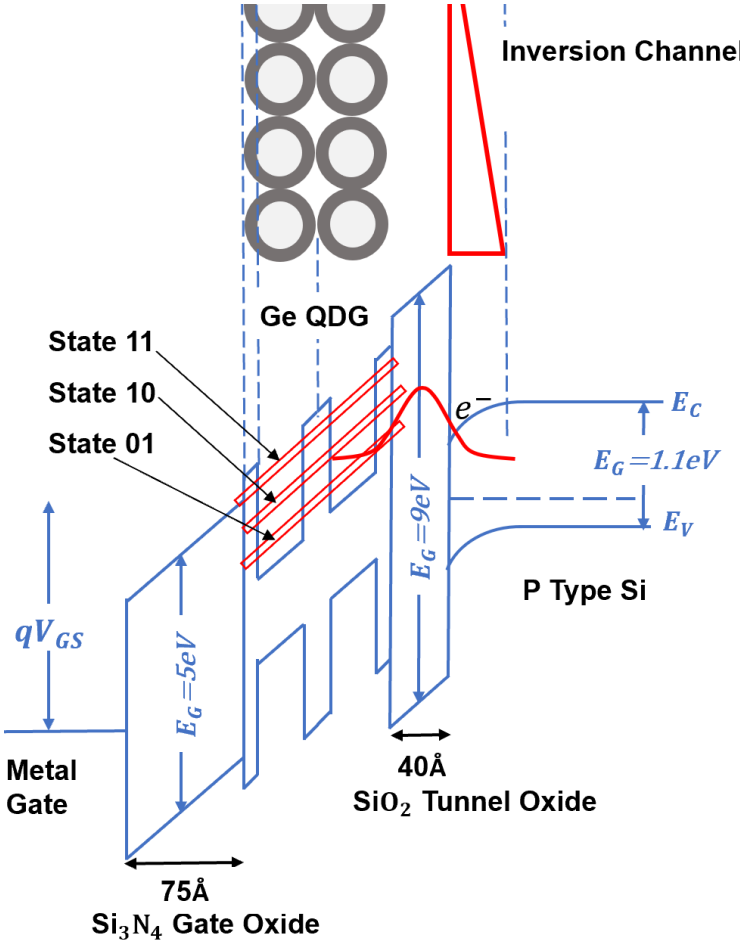


Figure 4.5: QDG Flash NVM Energy Band Diagram Writing Operation

Biasing the QDG NVM with a fixed V_{DS} and sweeping V_{GS} will yield an I_{DS} V_{GS} characteristics that has intermediate states between logic 0 and logic 1 as shown in figure 4.6. The fixed V_{DS} provide bias for electrons to travel through the inversion channel, as V_{GS} increases the resulting electric field promotes electrons to tunnel into the first miniband, as state 01. Once this miniband is full I_{DS} saturates for a while as V_{GS} increases. This occurs due to the fact that the electric field does not raise the electron's energy level high enough to tunnel into the second miniband and miniband 1 is already

fully populated. Once V_{GS} increases electron energy required to tunnel to second miniband, saturation occurs at state 10, for the same reason. As the V_{GS} sweep concludes state 11 miniband becomes populated. Figure 4.7 is experimental results reported by Lingalugari [11], of a similar device with mixed QD floating gate comprised of Ge and Si layers 1 and 2 respectively.

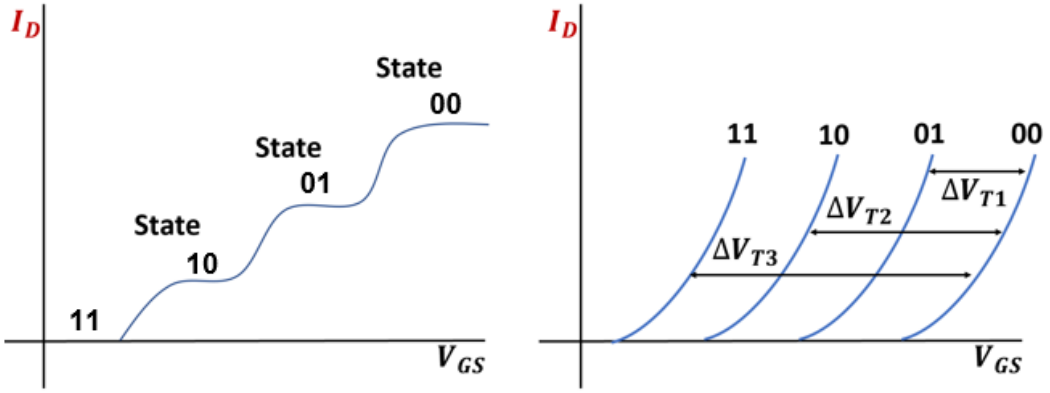


***Figure 4.6: Theoretical QDG Flash NVM I_{DS} V_{GS}* Characteristics and Multi ΔV_T**

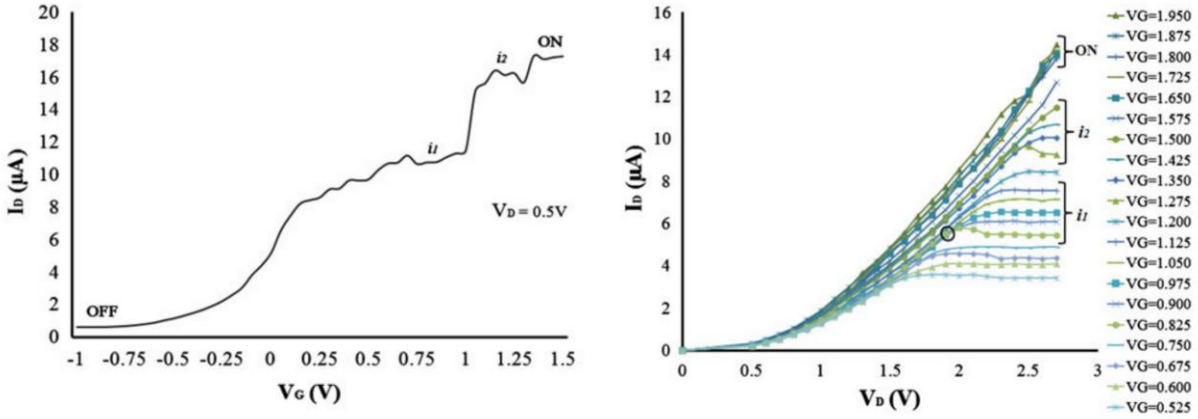


***Figure 4.7: Mixed Dot QDG FET I_{DS} V_{GS} (left), I_{DS} V_{DS} (right)* Characteristics [11]**

5.0 Fabrication of QDG NVM

Semiconductor fabrication requires high purity materials, controlled doping concentrations and quality oxide and layer growth, this is achieved by precise temperatures, times, chemical ratios and ultra-clean device surface throughout all processing steps from bare substrate to finished device. Wafer handling is also important, any spot an instrument touches, such as a tweezer, will leave damage and nonfunctioning devices. What separates the top manufactures in the semiconductor industry in terms of high yield, is well designed cleanrooms and procedures to minimize contamination, in a wide range of forms, and proper machinery and tooling that contacts the wafer to eliminate damage. In a university research setting the primary goal is to develop novel device technologies rather than achieve high yield since automation and cleanroom control is limited. To minimize damage fabrication was conducted under a fume hood with laminar air flow which along with basic lab coats, booties and gloves helped reduce particles. Also, wafers were handled by tweezers only and an effort made to only grab at the corners. The methods of fabrication utilized have be demonstrated effective by [11],[30],[31],[32],[33].

The QDGNVM device substrate is Silicon doped with Boron or P type wafer, which had a measured 1330Å thick field oxide due to exposure to the atmosphere. Filmetrics utilizing refractive index was the tool used to determine field oxide which becomes important when etching. The wafer was cleaned chemically, by heating three beakers with Trichloroethylene (TCE), Acetone and Methanol at 125ºC, and submerging the wafer in that order listed for 4 minutes in each chemical. TCE is a halocarbon that readily breaks down organic contaminates which is why it is widely used as a

degreaser. However, TCE leaves a film residue on wafers, which is why Acetone is used to remove it, and Methanol removes residues from Acetone. The wafer was blown dry with nitrogen gas.

The process of photolithography transfers a micro or nanometer geometric design onto the surface of the wafer, this is repeated for different geometries as a device is built layer by layer. Photoresist is spun onto the surface of the sample wafer at high rpm for uniform coating, 5000 rpm was used. Afterward it is placed on a hot plate to bake the photoresist onto the wafer and then cooled down to allow for a stable temperature to continue processing. Positive photoresist, such as S1813, is a chemical blend that when exposed to ultraviolet light (UV), undergoes a reaction. Photoresist in this state can be submerged in a developer solution, which dissolves all portion that were exposed to UV. Masks contain the designed geometries, comprised of a chrome pattern on a glass surface. The wafer is aligned underneath the mask, UV is applied above the mask, the chrome pattern blocks UV and UV in the transparent regions transmit trough to expose the photoresist below. The UV exposure time varies depending on strength of the UV source and was determined empirically when bulb was replaced and modified as the UV power reduces as the bulb ages. For processing it was determined S1813 should be exposed for 7 seconds and developed with 351 for 15 seconds to reproduce a high-fidelity mask pattern. The developer solution used is a 5:1 ratio of 125ml deionized (DI) water to 25ml 351. When developed, UV exposed portions are removed from the wafer leaving photoresist in the protected areas. Image reversal photoresist is also used in fabrication which works in the opposite manner, exposed UV remains, and UV shielded photoresist is developed away using different developer.

A mounting wafer is used to seal the back side of the sample wafer, to protect against unwanted phosphorous diffusion and or QD self-assembly on the bottom of the substrate in subsequent steps following photolithography. This prevents the formation of a parasitic PN junction on the substrate. A wafer larger than the sample wafer was rinsed with TCE, acetone and methanol and blow dried with nitrogen. It was loaded into a photoresist spinner, centered on a larger vacuum chuck, coated with S1813 positive photoresist and spun at 1000 rpm for 10 seconds which yields approximately 1.2µm coating. The sample wafer was pressed into the tacky photoresist layer, placed on a hot plate at 125ºC for 15 minutes to strengthen and air cooled for 2 minutes. The mounting wafer with the sample wafer adhered to its surface was reloaded and centered into the spinner, coated with S1813, spun at 5000 rpm for 30 seconds, baked on the hot plate and cooled resulting in a uniform coating, approximately 1.2µm thick, of photoresist covering the entire sample.

The “Mask 3, S/D Inside Well” was cleaned by rinsing with TCE / acetone / methanol, dried with nitrogen and loaded into the mask aligner equipment with the chrome pattern facing down. This mask opens source and drain regions and defines device location serving as a reference point for subsequent mask alignments. Mask 3 is roughly aligned such that the pattern is perpendicular to the vision field of the microscope. The mounting/sample wafer is loaded under mask 3 and elevated in height until it is almost touching the mask to ensure UV is focused where intended. The mask aligner equipment is cycled for 7 second UV exposure, the wafer is removed and developed in solution, outlined previously, for 15 seconds and the resulting pattern is inspected for quality under microscope. Ambient light can inadvertently develop photoresist, so the lab’s light was

switched to a yellow light during processing and the microscope light was reduced to a minimum for viewing. The wafer was rinsed in DI water, dried with nitrogen and then post baked for 10 minutes at 125ºC and cooled for 2 minutes to harden the remaining photoresist. Figure 5.1 is the process steps described so far in section 5.0.

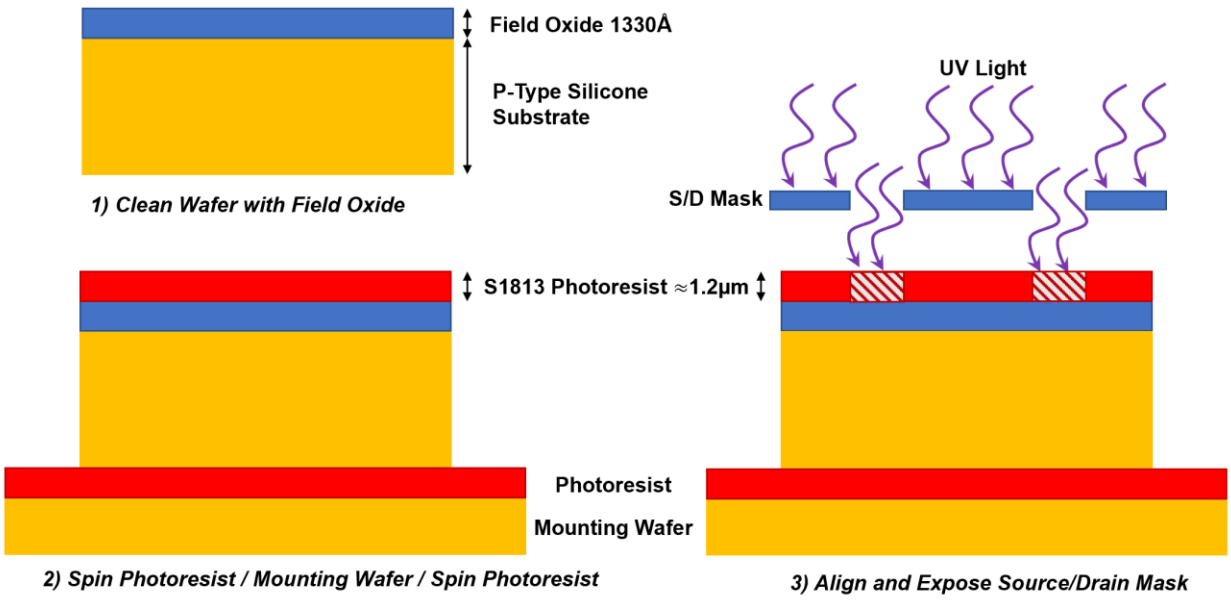


Figure 5.1: Fabrication Process File Steps 1, 2 and 3

Hardened photoresist is robust when subjected to the following processes not being impacted, while the developed sections experience chemical processing. The only areas not coated with photoresist are the source drain regions, covered with the 1330Å field oxide. Buffered Oxide Etch (BOE) removes SiO_2 at an etch rate dependent on BOE ratio per manufacture, see figure 5.2. [34] The BOE ratio 10:1 of NH4F: HF was mixed for etching which results in a 600Å / minute etch rate, therefore removal of field oxide is 2 minutes and 13 seconds. To accomplish this the wafer is rinsed in DI water, etched to remove field oxide and rinsed in a fresh beaker of DI water. The mounting/sample wafer was placed in a 125ºC acetone beaker to dissolve the photoresist until the two are

separated. The sample wafer was moved to a separate clean acetone beaker and soaked at 125ºC for an additional 15 minutes to remove all photoresist.

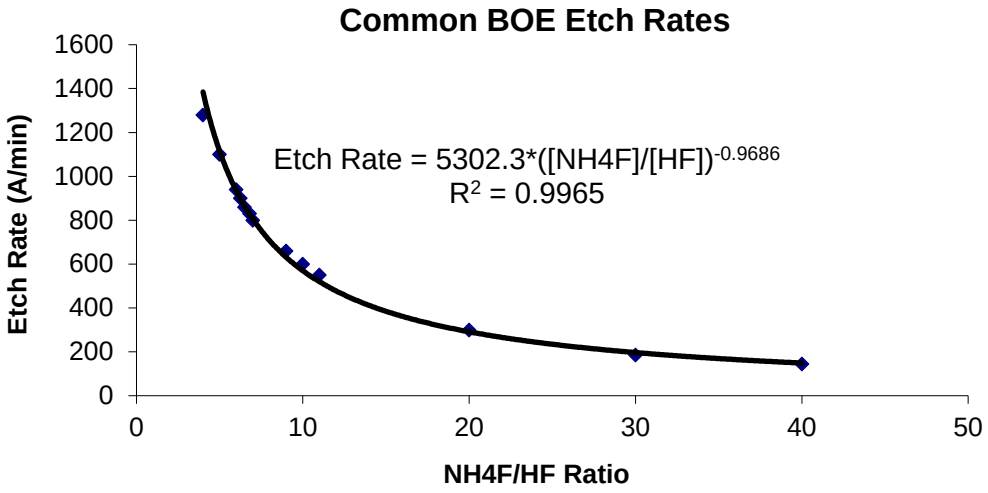


Figure 5.2: BOE Etch Rate [34]

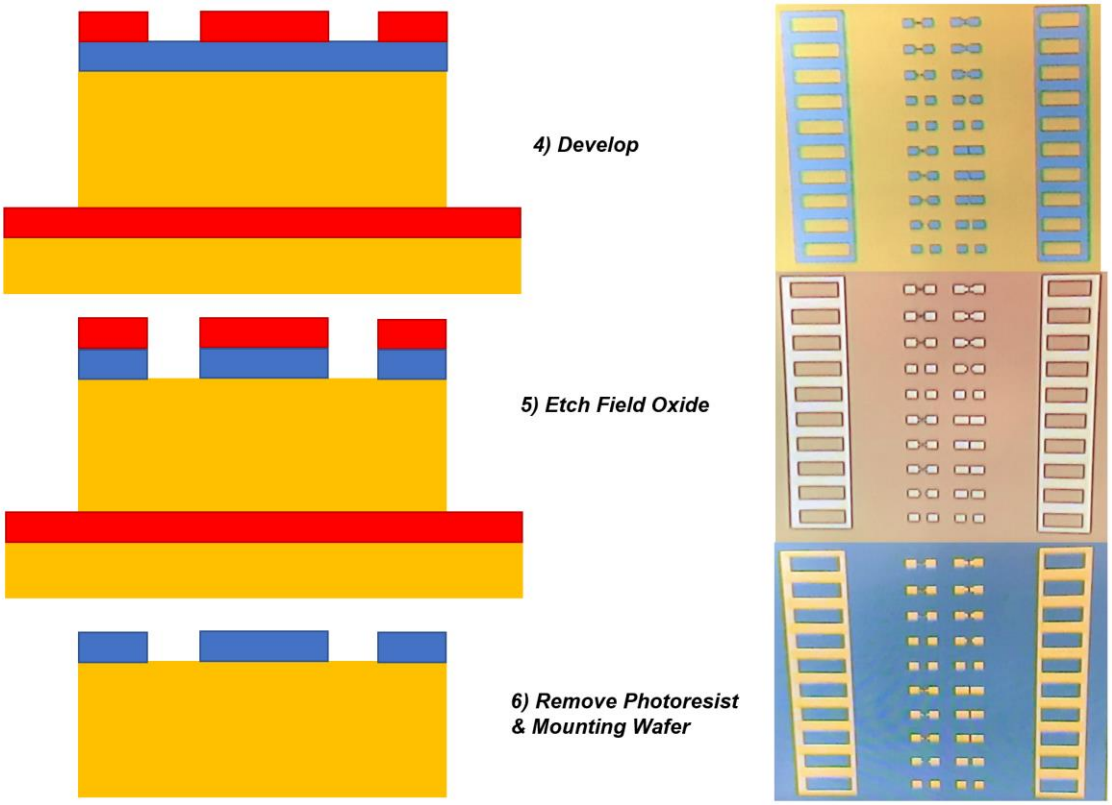


Figure 5.3: Fabrication Process File Steps 4, 5 and 6

The resulting condition of the sample wafer has the field oxide covering the entire wafer, except the small squares of the source drain regions which are unprotected revealing the P type substrate as shown in step 6 of figure 5.3 above. The sample was quickly loaded, (to avoid oxide growth on the substrate due to atmosphere), into the hot walled physical vapor deposition (PVD) chamber for phosphorous diffusion, see figure 5.4. PVD is a process of combining semiconductor materials, the chamber consists of a glass tube that is sealed and pumped to a vacuum, a surrounding furnace heats the chamber to 1000ºC. At these temperatures, the lattice of the semiconductor materials vibrates rapidly at the atomic level, interstitials are created, leaving voids in the lattice for atomized dopants to populate. Inert nitrogen gas is always flowed in the chamber to facilitate the migration of atomized dopants from the source disk to the wafer surface. The wafer is loaded onto the center groove of a quartz boat which stands it on edge, pure phosphorus source disks are loaded on either side of the wafer in adjacent slots.

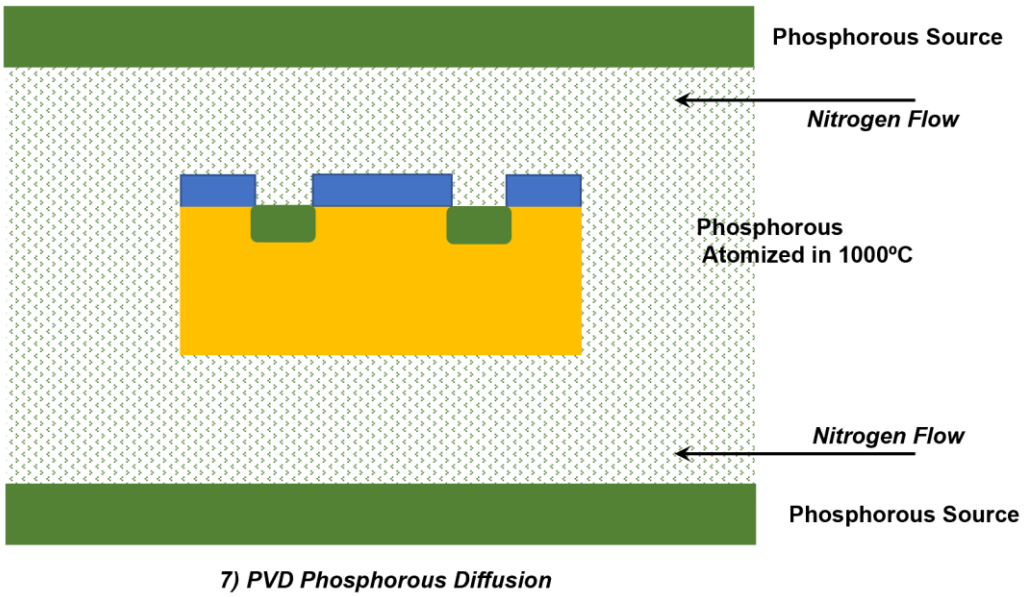


Figure 5.4: Fabrication Process File Step 7

The boat is loaded in the mouth of the chamber for 2 minutes to allow for a gradual heating to occur and avoid thermal shock, after which it is moved to the center of the chamber for 15 minutes of diffusion, after which it is gradually cooled in the chamber mouth for 2 minutes before removing. While in the center of the chamber the source disks emit a vaporized cloud of phosphorus, the flowing nitrogen deposits this vapor onto the exposed surface of the source drain regions, as the time progresses this pre-deposition layer is driven into the substrate resulting in an N type doped source drain.

The sample and mounting wafers were cleaned, re-mounted, and coated with photoresist as outlined previously. The next mask exposes the gate region of the sample while preserving source drain and the rest of the wafer with photoresist. This mask has an array of various gate geometries, all of which must be aligned to the array of source drain regions such that the gate region is centered in between with slight overlap, which is a difficult mask to align for all transistors in the array. The gate mask was cleaned and aligned to source drain, exposed, developed and post baked to prepare for etching. Figure 5.5 show the steps for gate mask alignment.

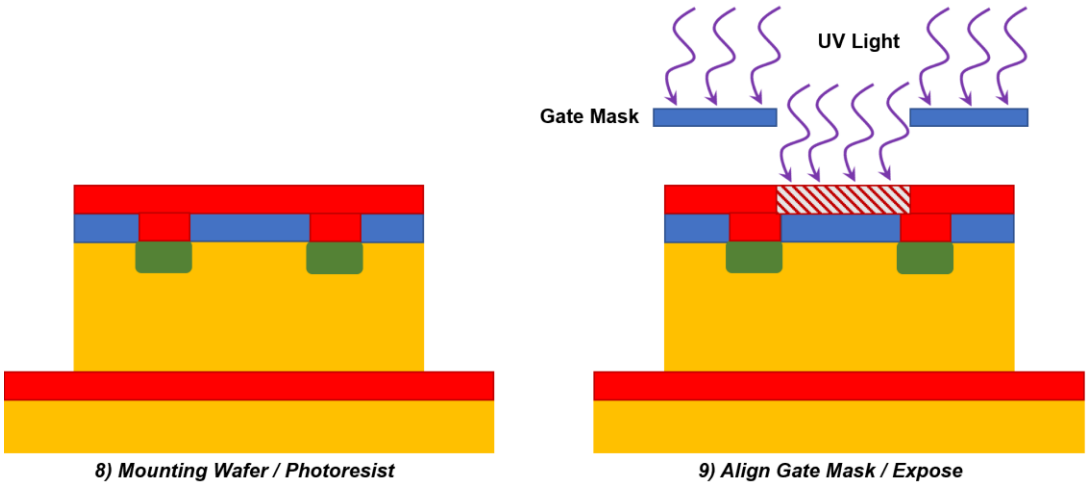


Figure 5.5: Fabrication Process File Steps 8 and 9

The field oxide being 1330 Å was etched for 2 minutes 13 second exposing the P type substrate in the gate region. Figure 5.6 show the steps for exposing the gate the photographs of the device shown to the right link to the process files. Typically, additional cleaning steps are performed by growing a sacrificial wet oxide layer encapsulating particles and then etching it off to ensure all contamination is removed. A piranha solution (1:1 of H2SO4:H_2O_2) is an aggressive wet oxidizer that quickly grows this sacrificial layer to be etched. The wet oxide clean was not conducted for this sample.

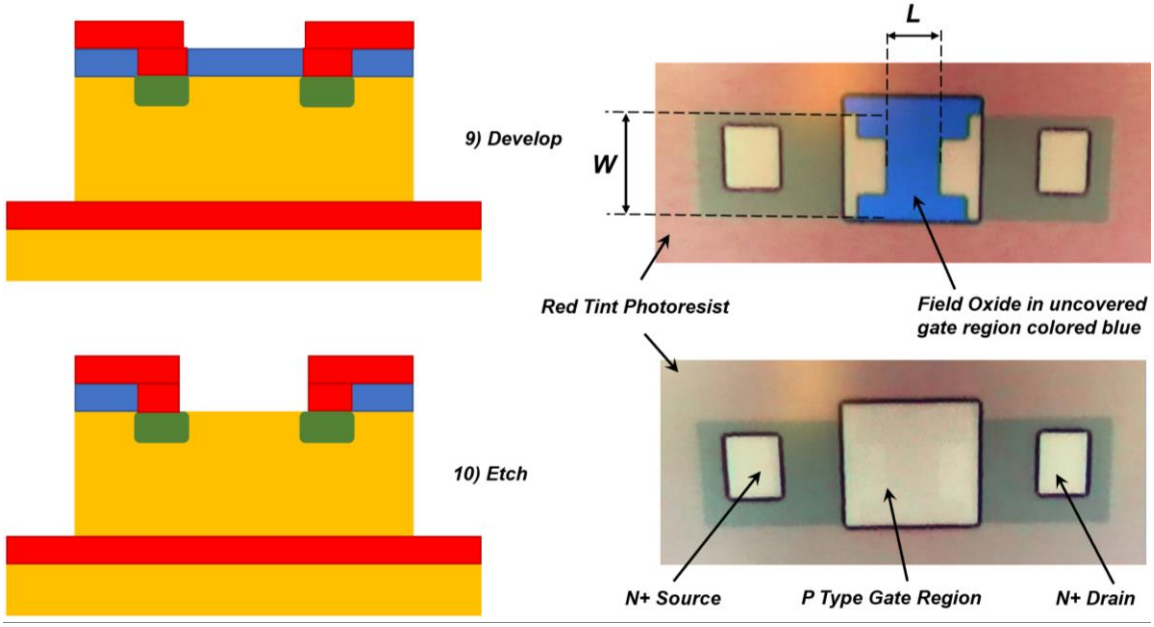
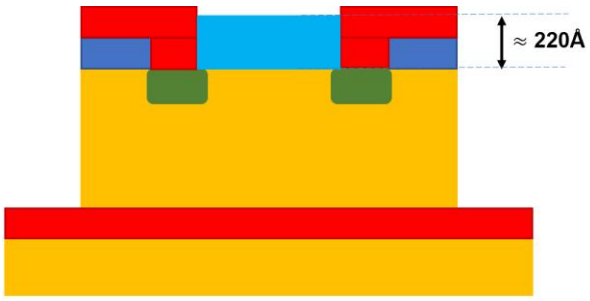


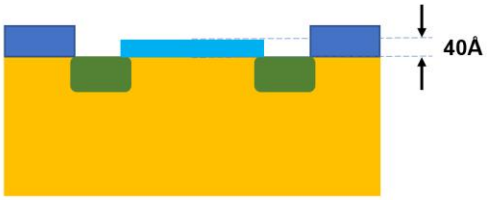
Figure 5.6: Fabrication Process File Steps 9 and 10

The sample wafer was soaked in acetone until it was separated from mounting wafer and then moved to a clean acetone soak to remove all photoresist. High quality tunnel oxide is required; therefore, a dry oxidation Chemical Vapor Deposition (CVD) furnace was used for SiO_2 growth. This design requires direct tunneling which occurs in barriers between 60Å and 20Å thick, the thin end of that range approaches the limit of being able to stop electrons from tunneling at all, even in unbiased structures. During

the first fabrication run the tunnel oxide was grown, from P type substrate up to 20Å which resulted in a leaky gate, on one device all drain current leaked to the gate during the gate voltage sweep. Either 20Å was too thin or the method of growth inaccurate. The second sample fabrication run increased the tunnel oxide to 40Å and altered the method of growth. Since the growth rate of dry oxide growth for this particular furnace is not linear, it was chosen to grow more oxide than needed, then measuring, and etching down to 40Å at a known rate. The CVD dry oxide furnace is comprised of a glass tube chamber, under vacuum, with nitrogen flowing as it is heated to 1000ºC. The wafer loaded in a quartz boat was placed at the mouth of the chamber, nitrogen gas was turned off and oxygen flow turned on, the boat sat in the mouth for 15 minutes allowing oxygen to completely fill the chamber. The boat was moved further in the mouth but outside of the high temperature zone to thermally adjust, before being positioned at the center of the furnace for 10 minutes. Before removing the wafer from the chamber, it spent an additional 2 minutes in the mouth reducing heat at a slower rate to avoid thermal shock. It was empirically found with Filmetrics that these locations and times produced a 215Å to 220Å thick SiO_2. At the 10:1 BOE ratio 18 seconds were used to etch the final tunnel oxide to approximately 40Å as designed. Figure 5.7 shows the process file steps and photograph of resulting tunnel oxide.



11) CVD Dry Oxide Growth



12) Etch Tunnel Oxide
Remove Mounting Wafer & Photoresist

Field Oxide surrounding device blue color

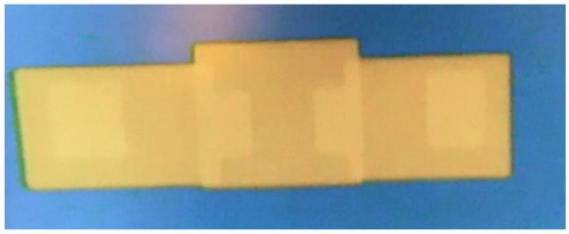


Figure 5.7: Fabrication Process File Steps 11 and 12

Fabrication of Ge QD clad with GeO_2, starts with pure germanium powder which is milled with glass ball bearings into an ultra-fine powder. This milled Ge powder is mixed with ethel alcohol and benzenol peroxide and sonicated for at least a week, this batch was under ultrasonic vibration for about a month. This process separates the powder into individual particles and grows the oxide cladding. This solution is extracted and centrifuged in test tubes for these sequential times and speeds: 30 minutes at 3k rpm, 30 minutes at 6k rpm, 30 minutes at 9k rpm and 60 minutes at 13k rpm. Larger Ge particles in the solution precipitates leaving a clear colloidal solution on top consisting of 4nm diameter Ge QD. A solution was made for self-assembly by extracting 5kμL of the QD colloidal solution and mixing with 4.9kμL of ethel alcohol and 100μL of BOE.

The cladding shells of the QD are negatively charged and as such they will electrostatically be attracted to positively charged material, therefore QD will self-assembly over the tunnel oxide due to the P type channel underneath.

The sample wafer is cleaned with TCE, acetone and methanol, rinsed in DI water, rinsed in methanol and dried with nitrogen before being submerged in the QD solution for three minutes at room temperature. After QD solution soak, the sample wafer is removed, rinsed with methanol and dried with nitrogen. This procedure has been empirically shown to produce two layers of QD. To strengthen the QD floating gate layers the sample wafer was annealed by placing in a furnace, ramping temperature to 350ºC and maintaining this temperature for five minutes. The furnace chamber had a constant flow of argon throughout annealing. The sample wafer was brought to another university laboratory to grow the 75Å of silicon nitride, Si_3N_4, control gate insulation layer over the QD floating gate region using Plasma Enhanced CVD. Figure 5.8 shows the QDG and control gate oxide growth.

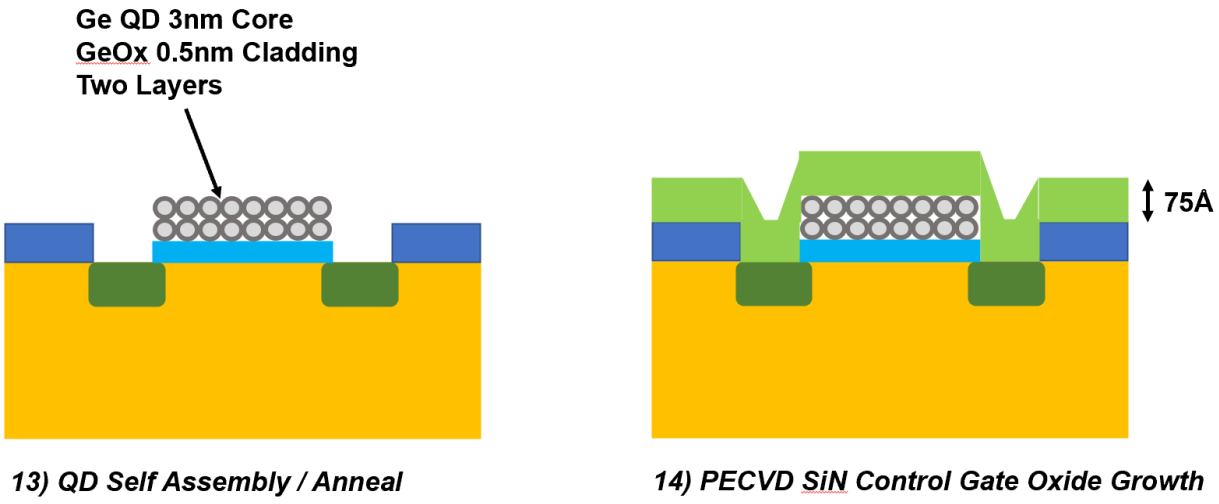


Figure 5.8: Fabrication Process File Steps 13 and 14

The sample wafer was cleaned, spin coated with photoresist and hardened on the hotplate and the source drain contact mask was cleaned, dried and loaded into the mask aligner. Alignment resulted in the contact openings centered in the source drain regions, exposure, development, inspection and bake hardening was executed. The exposed

Si_3N_4 control gate oxide, plus a negligible amount of oxide that grew after phosphorous diffusion was etched with BOE for 8 seconds to remove 75Å at the 600Å per minute etch rate down to the substrate. A 10-minute acetone soak removed all photoresist. The opening of source and drain contacts is detailed in figure 5.9.

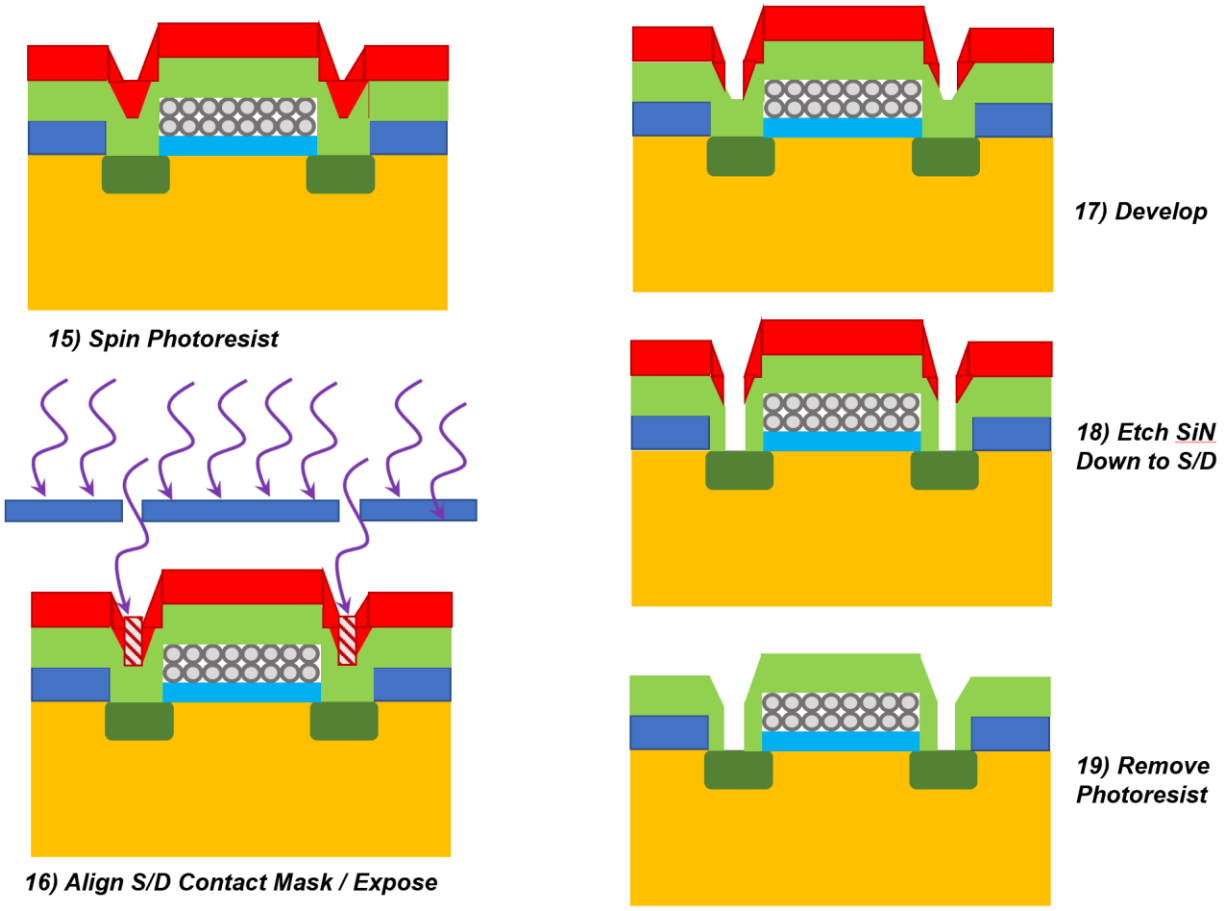


Figure 5.9: Fabrication Process File Steps 15, 16, 17, 18 and 19

Aluminum (Al) metallization sputtering deposited metal for source, drain and gate contacts. The bell jar was pumped down in pressure, after 2 hours the ionization gauge read $2x10^{-5}$ Torr, liquid nitrogen was added to reduce to $2x10^{-6}$ Torr or lower. The sample wafer was loaded, and the roughing valve opened, until pressure reached 10-20 Torr at which point the roughing valve was closed. The foreline and gate valve were opened until the pressure reduced to $2x10^{-6}$ Torr with the addition of liquid nitrogen. A tungsten filament

with high current of 30 amps, passed through Al pellets, causing the aluminum to atomize. When the pressure reached $1x10^{-5}$ Torr the shutter was opened to allow the Al cloud to exit at which point the current was increased through the tungsten filament and pressure increased to $2x10^{-5}$ Torr. The Al was deposited on the sample wafer growing a layer of metal, an Infacon meter was monitored until 1500Å metal thickness was achieved. Previous fabrication of trying a thin Al layer to 600Å resulted in poor contact with terminals and undesired Al flaking during metal etch and testing. The metalized sample wafer was cleaned, spun with photoresist and the last interconnect mask was cleaned, aligned, exposed and developed. The exposed Al is etched with an Aluminum Etchant followed by the removal of the photoresist by soaking in PG Remover at 80ºC for 2 hours. Figures 5.10 and 5.11 show the metal etching and resulting finished QDG Flash NVM device structure.

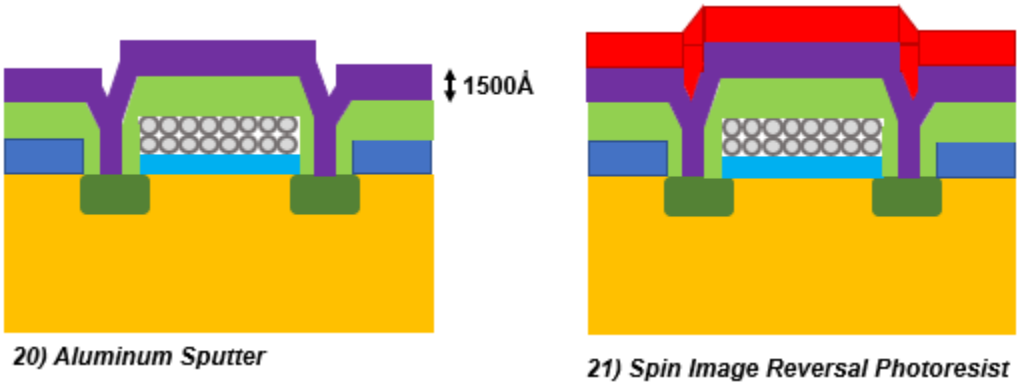


Figure 5.10: Fabrication Process File Steps 20, 21,

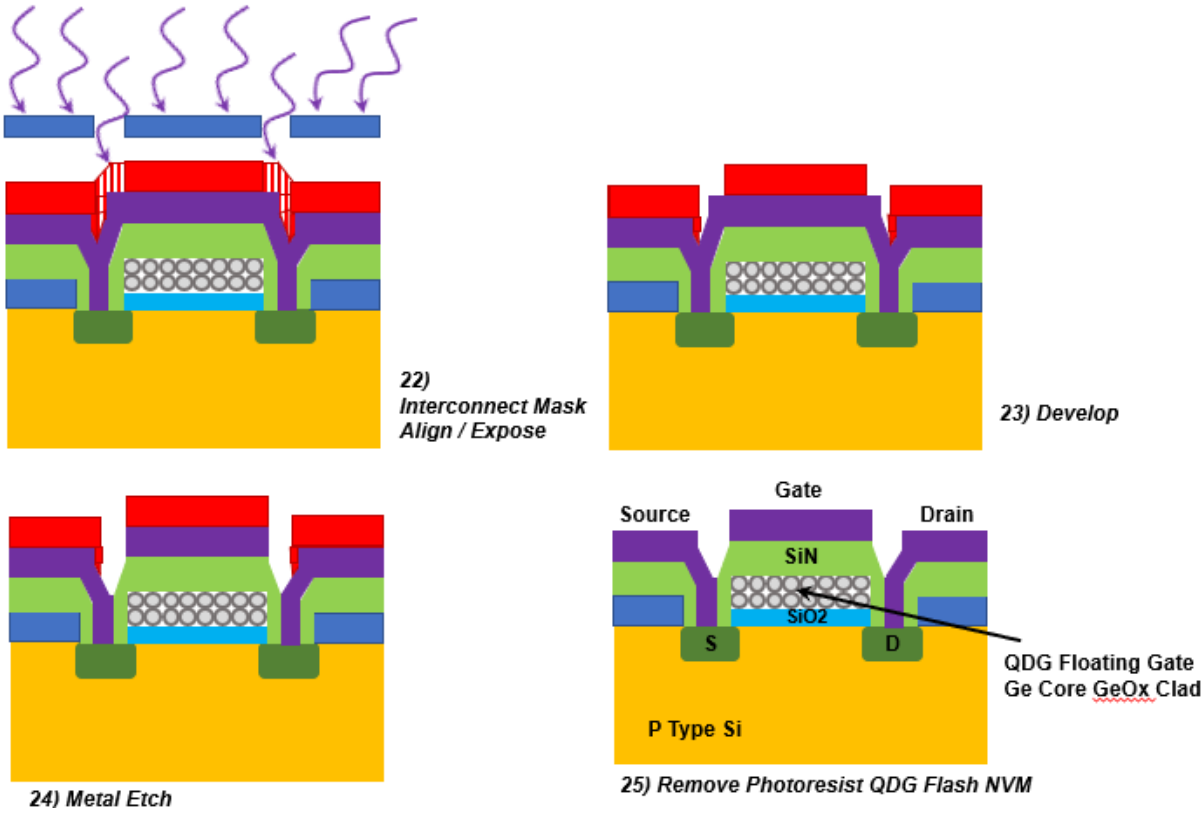


Figure 5.11: Fabrication Process File Steps 22, 23, 24 and 25

6.0 Results

The testing of the device outlined in the processing step in the section 5.0 of this thesis had two versions fabricated. The first round the tunnel oxide was grown to 20Å thick. This was either not controlled well enough during fabrication and was too thin for direct tunneling or the growth was not uniform. During testing of this device, the current was leaking to the gate for all NVM cells on the wafer, for one of specific devices tested, it was observed the $I_D = I_G$ indicating no insulation between channel and gate. It was observed that the write pulse impacted the devices, however, it is unclear if charge was being stored in the QD since the gate leakage was too great to interpret although other device test, to included herein, resulted in write pulse alterations to the $I_D\ V_G$ characteristics. The $I_D\ V_G$ characteristics as shown in figure 6.1 are the testing results of sample 3 before and after a 10V write pulse. Other devices showed saturating currents during various sweep patterns; however, the gate leakage was too high and this device did not produce results.

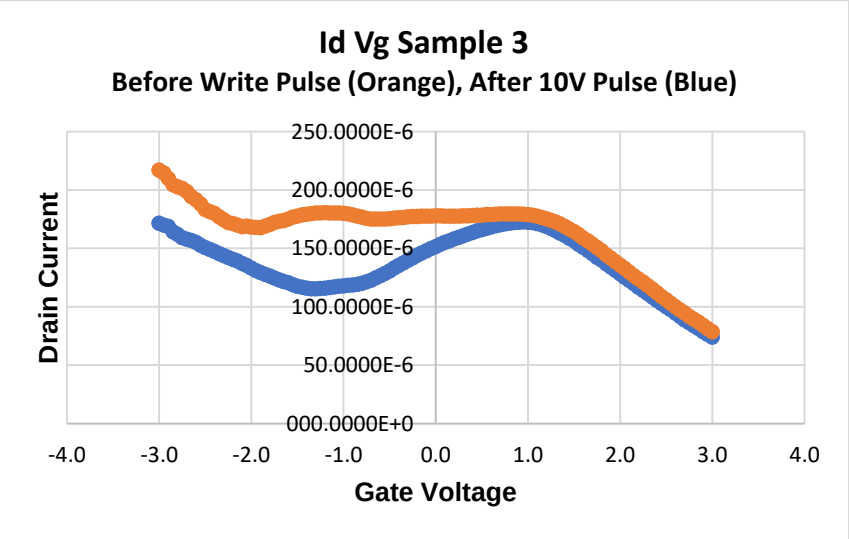


Figure 6.1: QD Storing Charge Impact Observed, Sample 3 Before (orange) and After 10V Write Pulse (blue)

As the gate voltage increased the drain current decreased since it was leaking to the gate which was observed on measurement equipment but not shown in graph. One

device from this sample showed the drain current equaled the gate current indicating 100% leakage.

To correct fabrication issues the mounting wafer was implemented during fabrication as outlined by process steps to protect the P type substrate from phosphorus and QD self-assembly. This measure eliminates additional, parasitic PN junctions from forming. The tunnel oxide was increased to 40Å to increase the barrier. Also as indicated the oxide was grown thicker and etched back to 40Å to have increased control over the growth to ensure uniform layer as opposed to growing from 0 to 20Å on the previous failed attempt. Due to the COVID-19 pandemic fabrication and research suspended work and the Si_3N_4 gate control oxide, metallization and testing are pushed to future work.

A previous device that was fabricated by N. R. Butterfield *et al.* 2016 for ECE6244, had promising results and also requires more testing. This QDG NVM also has a Quantum Dot Channel (QDC) between source and drain. The P type substrate and N+ source drain have identical processing as outlined. The QDG and QDC are comprised of Si core SiO_2 clad dots, hafnium oxide (HfO_2) was used as a tunnel and control gate barrier. HfO_2 has a band gap of 5.7eV and is used due to the high k dielectric constant which alters C_{ox} allowing for thinner layers of insulation compared to SiO_2. Nitride was also used to protect source and drain during fabrication. A thick layer of wet SiO_2 was grown and etched above the channel to remove a thin layer of the substrate, allowing for the self-assembly of the QDC to occur between so. Figure 6.2 shows the layers of this device.

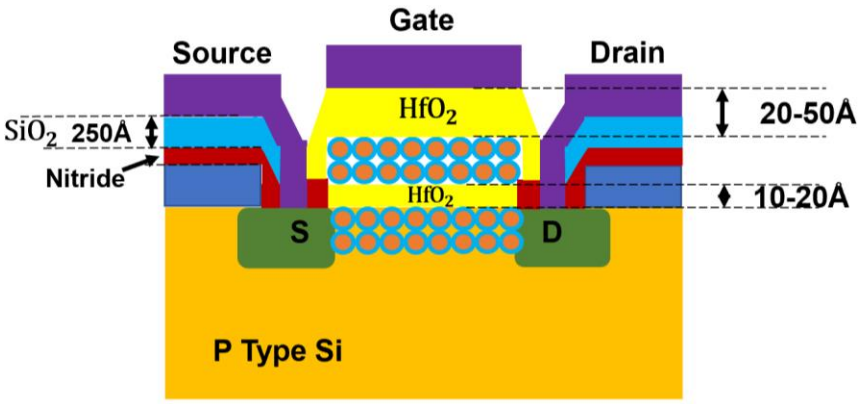


Figure 6.2: Fabricated QDC-QDG Flash NVM

The device fabricated and schematically show in figure 6.2 was tested for various fixed V_D and the V_{GS} swept from 0V to 6V. It was demonstrated that with a fixed drain voltage of 3V, multistate of the device was observed, as shown by the data collected in figure 6.3. These I_D V_G Characteristics shows three traces after gate pulses applied and identical biasing with two distinct I_D V_G characteristic shifts are indicated by the superimposed blue arrows in figure 6.3.

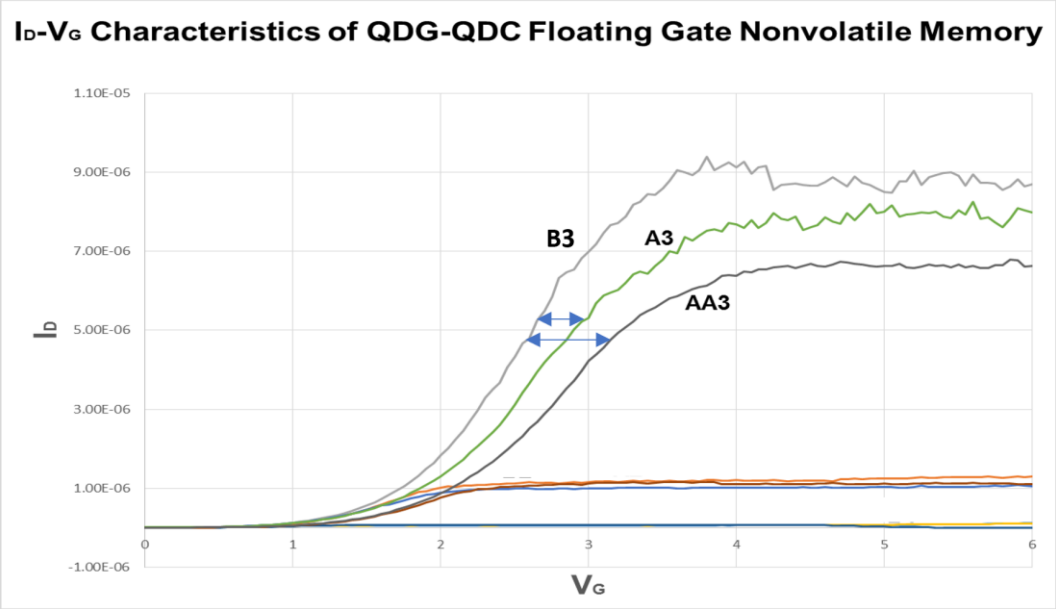


***Figure 6.3:* I_D V_G Characteristics, B3 (No Pulse), A3 (After 12V 100µs Write Pulse), AA3 (After 15V 100µs Write Pulse)**

Trace B3 is I_D V_G characteristic for no voltage pulse and a fixed drain voltage of 3V. After a 12V for 100µs pulse to the gate, the same gate sweep was conducted producing A3 trace at a fixed V_D=*3V.* Then 15V for 100µs pulse to the gate, the same gate sweep was conducted producing AA3 trace at a fixed V_D=*3V.* There are shifts between all three traces due to the minibands in the QDSL, due to the QDC, however, there is no shift in threshold voltage. This makes using a read pulse not possible to implement since there is not enough separation. For example, a 3V read pulse would have a drain current of approximately $4x10^{-6}$A for trace AA3, $5.2x10^{-6}$A for trace A3 and $7x10^{-6}$A for trace B3. This low level of separation in currents makes this device difficult to distinguish states without a clear shift in threshold voltage. The testing conducted was sweeping the gate voltage for three fixed drain voltages. This was repeated for three trials: 1) before any write pulse, 2) after 12V for 100µs and then 3) after 15V for 100µs.

7.0 Future Work

To fix the leaking gate a higher quality and thicker tunnel oxide will need to be fabricated. The nitride layer in the QDG-QDC device acted as a protection layer during fabrication and showed promise in protecting source and drain. The mounting wafer should also be used in all future work.

The device outlined in the process files needs to resume fabrication upon the research moratorium being lifted. Purposed testing is as follows:

1. With drain and source grounded, erase QDG to ensure it has no charge from fabrication which is also industry best practice by pulsing V_{GS} -10V for 10ms.
2. Sweep V_{GS} for various fixed V_D to empirically determine best characteristics for further testing also sweep V_D for various fixed V_{GS} in the same manner.
3. Implement more write pulses 10V to 20V in increments of 2V for 100µs. It is likely there are many more states that are not being tested under current testing methods.
4. Erase device with V_{GS} -10V for 10ms
5. Try and program specific minibands by pulsing from 20V to 10V decreasing by 2V and erasing between trials. This will be able to test if an intermediate state can be selected without previous write pulses prefilling lower minibands.

Regarding the QDG-QDC Flash NVM Device fabricated previously and included in the results section, more testing is needed. It is likely there are more states able to be programed by following the purposed testing outlined above in future work to investigate if a significant voltage threshold shift can be measured.

8.0 Conclusion

QDG Flash NVM devices requires more work and testing to corroborate test results within this lab and from others in academia to prove repeatable, functional and scalable results. Future work should include efforts to increase drain current saturation during multistates observed for I_D V_G characteristics and increasing the shift in threshold voltages between states, both of which will make the intermediate states more viable. A more thorough investigation into revealing additional states as well as testing if a certain intermediate state can be specifically programmed should also be perused.

Moore's Law can also be applied to other technological arcs throughout history, starting with an invention, improvement of that technology, plateau and obsolescence's. Replacement technologies overlap and then repeat this exponential progression. This is the nature of advancement; complexity, accuracy and reliability results in growth in all three. The first computers filled rooms and consumed large amounts of power and time to compute mathematics that are easily solved with paper and pencil. All the computers used in the Apollo 11 mission that put a man on the moon are dwarfed by the computing power in the current smart phone. This increased computing power allows increasingly difficult problems to be solved in shorter time periods, with less power consumption and manufacturing costs. Moore's Law has proven to be an accurate forecast of transistor performance and size. Currently in the year 2020 the end of the law for the conventional transistor has come and advancement is plateauing as the FET structure is altered to extend the curve.

Information is a crucial component of nature and technology and will always be essential, requiring transmission, storage and processing. Therefore, memory cell

technology will be required. The current period of technological advancement in computing as of 2020 is bridging the gap between the transistor-based memories of the past and technologies of the future, through altering current architectures and developing new ones. Binary states have endured since punch cards and will be utilized into the future. However, multistate devices such as the QDG Flash NVM, QD SRAM and QD DRAM as well as quantum computing's use of logic 1, 0 and Qbit may be pointing the path into the future. Regardless of the next technology, or group of technologies, to replace transistor-based memory devices, they likely are in the research and development phase now and will likely follow the arc of a new Moore's Law.

9.0 References:

[1] T.N Theis, H.S.P Wong. The End of Moore's Law A New Beginning For Information Technology. Copublished by the *IEE CS* and the *AIP.* 1521-9615/16, 2016. https://e3s-center.berkeley.edu/wp-content/uploads/2019/06/2017_The-End-of-Moore%E2%80%99s-Law-A-New-Beginning-for-Information-Technology.pdf

[2] J. Cooke, Micron Technology Inc. NAND Flash:101 An Introduction to NAND Flash and How to Design It In to Your Next Product. *PDF:09005aef8245f460 / Source: 09005aef824f,* tn2919_nand_101.fm -RevB 4/10 EN, 2007 https://www.techonline.com/electrical-engineers/education-training/tech-papers/4126198/NAND-101-An-Introduction-to-NAND-Flash-and-How-to-Design-It-into-Your-Next-Product

[3] S. Inderjt. Electronic Devices 2016. *Weebly*. 2016 https://inderjitsingh87.weebly.com/electronic-devices-2016.html

[4] H-W. Wook, C. Kang, Y-S. Kim. A Novel Structured Polysilicon Thin Film Transistor that Increases the On/Off Current Ratio. *Semiconductor Science and Technology.* Volume 18 Number 9, 2003 https://iopscience.iop.org/article/10.1088/0268-1242/18/9/306

[5] J. Ranuarez, M.J. Deen, C-H. Chen. A Review of Gate Tunneling Current in MOS Devices. *Microelectronics Reliability.* Volume 46, Issue 12 Pages 1939-1956, 2006. https://www.sciencedirect.com/science/article/abs/pii/S0026271406000205

[6] B. Obi Tayo. Finite Difference Solution of the Schrodinger Equation. *www.medium.com/modern-physics.* 2019 https://medium.com/modern-physics/finite-difference-solution-of-the-schrodinger-equation-c49039d161a8

[7] M.R. Zakaria, M.N. Hashim, U. Hashim, R.M. Ayub, T. Adam, A.W. Al-Mufii. An Overview and Simulation Study of Conventional Flash Memory Floating Gate Device Using Concept F-N Tunneling Mechanism.*Fifth International Conference On Intelligent System, Modeling and Simulation IEEE.* DOI 10.1109/ISMS.2014.142, 2014 https://ijssst.info/Vol-15/No-3/data/3857a775.pdf

[8] A. Padilla. Advanced Structures and New Detection Methods for Future High Density Non-Volatile Memory Technologies. *University of California at Berkely.* Doctorate of Philosophy Dissertation, 2007. https://people.eecs.berkeley.edu/~tking/theses/apadilla.pdf

[9] S.B Desai, S.R. Madhvapathy, A.B. Sachid, J.P. Llinas, Q. Wang, G.H. Ahn. MoS2 Transistors with 1 Nanometer Gate Length. *Science,* Vol. 354, Issue 6308 pp. 99-102 https://science.sciencemag.org/content/354/6308/99

[10] H.J.M. Veendrick. Nanometer CMOS ICs From Basics to ASICS Second Edition. *Springer.* ISBN 978-3-319-47595-0, 2017
https://link.springer.com/book/10.1007/978-3-319-47597-4#about

[11] M. Lingalugari, K. Baskar, P.Y. Chan, P. Dufilie, E. Suarez, J. Chandy, E. Heller, F.C. Jain. Novel Multistate Quantum Dot Gate FETs Using SiO2 and Lattice-Matched ZnS-ZnMgS-Zns as Gate Insulator. *Journal of Electronic Materials,* Vol 42 NO.11, 2013
https://link.springer.com/article/10.1007%2Fs11664-013-2696-7

[12] Common BOE Etch Rate Graph .xls
https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Faskavantor.secure.force.com%2Fservlet%2FfileField%3Fid%3D0BEG0000000TSTP

[13] T. Ytterdal, Y. Cheng and T.A. Fjeldly. Device Modeling for Analor and RF CMOS Circuit Design. *John Wiley & Sons, Ltd*. Chapter ISBN 0-471-49869-6, 2003

[14] P. Pavan, R. Bez, P. Olivo, E. Zanoni. Flash Memory Cells An Overview. *Proceedings of the IEEE.* Vol. 85, NO. 8, 1997
https://www.semanticscholar.org/paper/Flash-memory-cells-an-overview-Pavan-Bez/49f39786b87dddbf088ae202f2bdad46668387e3

[15] F.C. Jain. ECE 6244 Nanotechnology II Laboratory (Spring 2016) Design and Fabrication of Nanodevices and Systems. *University of Connecticut.* 2016

[16] F.C. Jain. ECE 4242 Micro/Opto-Electronic Devices and Circuits Fabrication Laboratory Instruction Set (Fall 2014). *University of Connecticut.* 2014

[17] F.C. Jain. ECE 4211 Semiconductor Devices & Nanostructures Spring 2014. *University of Connecticut,* Vol 1 and Vol 2, 2014

[18] C.C. Hu Modern Semiconductor Devices for Integrated Circuits. *Prentice Hall.* ISBN-13: 978-0-13-608525-6, Chapters 4, 5 and 6, 2010

[19] P. Mirdha. Optimizing Fabrication and Modeling of Quantum Dot Superlattice for FETs and Nonvolatile Memories. *University of Connecticut*. Doctor Dissertations, 1686, 2017
https://opencommons.uconn.edu/cgi/viewcontent.cgi?article=7924&context=dissertations

[20] S. Karmakar. Design of Quaternary Logic Circuit Using Quantum Dot Gate-Quantum Dot Channel FET (QDG-QDCFET). *International Journal of Electronics*. 101:10, 1427-1442, 2014
https://www.tandfonline.com/doi/abs/10.1080/00207217.2013.854549?src=recsys&journalCode=tetn20

[21] P. Mirdha, B. Parthasarathy, J. Kondo, P-Y Chan, E. Heller, and F.C. Jain. An Investigation of Quantum Dot Super Lattice Use in Nonvolatile Memory and Transistors.

Journal of ELECTRONIC MATERIALS. Vol. 47, No. 2 page 1371-1382, 2018 https://link.springer.com/article/10.1007/s11664-017-5951-5

[22] J.J. Makwana, D.K. Schroder. A Nonvolatile Memory Overview. *www.aplawrence.com*. 2020. https://aplawrence.com/Makwana/nonvolmem.html

[23] R. Novotny, J. Kadlec and R. Kuchta. NAND Flash Memory Organization and *Operations. Information Technology & Software Engineering*. Vol. 5, Issue 1 page 1-8, 2015 https://www.longdom.org/open-access/nand-flash-memory-organization-and-operations-2165-7866-1000139.pdf

[24] M. Kang, J-S Yang, and IK-J Chang. Studying Trapped Tunneling-Electron Migration Due to Program and Erase Cycles in NAND Flash. *IEEE Electron Device Letters*. Vol. 37, No. 3 page 284-286, 2016. https://ieeexplore.ieee.org/abstract/document/7403909/similar#similar

[25] S.R. Kodigala. Verification of Fowler-Nordheim Electron Tunneling Mechanism in Ni/SiO_2/n-4H SiC and n^+ poly-Si/SiO_2/n-4H SiC MOS Devices by Different Models. *Physica B: Condensed Matter*. Volume 500, pages 35-43, 2016 https://www.sciencedirect.com/science/article/abs/pii/S0921452616302939

[26] K. Ramkumar *et al*., "A scalable, low voltage, low cost SONOS memory technology for embedded NVM applications," *2013 5th IEEE International Memory Workshop*, Monterey, CA, 2013, pp. 199-202, doi: 10.1109/IMW.2013.6582134. https://ieeexplore.ieee.org/document/6582134

[27] M-H. Jung. K-S Kim, G-H. Park and W-J Cho. Dependence of charge trapping and tunneling on the silicon nitride thickness for tunnel barrier engineered nonvolatile memory applications. *Applied Physics Letters,* Vol 94, Issue 5, 2009. https://aip.scitation.org/doi/abs/10.1063/1.3078279?journalCode=apl

[28] M. Cubukcu *et al*., "Ultra-Fast Perpendicular Spin–Orbit Torque MRAM," in *IEEE Transactions on Magnetics*, vol. 54, no. 4, pp. 1-4, April 2018, Art no. 9300204, doi: 10.1109/TMAG.2017.2772185. https://ieeexplore.ieee.org/document/8291048

[29] T. Inokuchi, H. Yoda, M. Shimizu, *et al.* "Improved Read Disturb and Write Error Rates in Voltage-Control Spintronic Memory (VoCSM) by Controlling Energy Barrier Height." *Applied Physics Letters,* Volume 110, Issue 25, 2017. https://ui.adsabs.harvard.edu/abs/2017ApPhL.110y2404I/abstract

[30] J. Kondo Quantum Dot Channel Field-Effect Transistors and Non-Volatile Memories: Fabrication and Simulation. *University of Connecticut*. Doctor Dissertations, 1455, 2017. https://opencommons.uconn.edu/dissertations/1455/

[31] R.S. Velampati. Quantum dot gate nonvolatile memory: Modeling and fabrication. *University of Connecticut*. Doctor Dissertations, 2007.

[32] F.C. Jain, S. Karmakar, P-Y. Chan, E. Suarez Quantum Dot Channel (QDC) Field-Effect Transistors (FETs) Using II–VI Barrier Layers. *Journal of Electronic Materials,* 41, 2775–2784, 2012. https://link.springer.com/article/10.1007/s11664-012-2161-z

[33] F.C. Jain, E. Suarez, M. Gogna, F. Alamoody, D. Butkiewicus, R. Hohner, T. Liaskas, S. Karmakar, P.-Y. Chan, B. Miller, J. Chandy & E. Heller. Novel Quantum Dot Gate FETs and Nonvolatile Memories Using Lattice-Matched II–VI Gate Insulators. *Journal of Electronic Materials* volume 38, pages1574–1578, 2009
https://link.springer.com/article/10.1007/s11664-009-0755-x

[34]
https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Faskavantor.secure.force.com%2Fservlet%2FfileField%3Fid%3D0BEG0000000TSTP